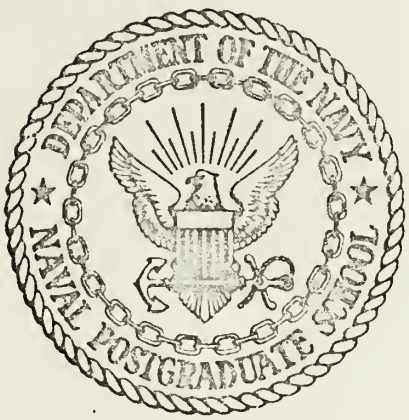
STOCHASTIC APPROXIMATION: A SURVEY

Thomas Keith Scheber

NAVAL POSTGRADUATE SCHOOL
Monterey, California



THESIS

STOCHASTIC APPROXIMATION: A SURVEY

by

Thomas Keith Scheber

Thesis Advisor: Bruno Shubert

March 1973

Stochastic Approximation: A Survey

by

Thomas Keith Scheber
Lieutenant, United States Navy
B.S., United States Naval Academy, 1967

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1973

ABSTRACT

Since 1951 when Robbins & Monro's pioneering paper on stochastic approximation was published, many articles have appeared dealing with extensions, modifications, methods and applications of stochastic approximation. While the concepts involved are relatively simple, but mathematically difficult, the information concerning specific results has been widely scattered and difficult to collect for the interested researcher. This paper will attempt to discuss the major results and will provide the necessary references to direct the user to more specific findings.

TABLE OF CONTENTS

I. INTRODUCTION

In many areas of analysis in bioassy, sensitivity testing or learning we are concerned with a level of output, Y, given a certain level of some input, x. For each given level of x, the resultant output is not deterministic but has some underlying probability distribution, $F(Y|X)$. Hence it is then common to refer to the response function of x, denoted $M(x)$, as the expected value of Y given x. (I.e., $M(x) = \int_{-\infty}^{\infty} Y(x)dF(Y|x) = E[Y|x]$.)

In usual analysis of the response function, $M(x)$, it is assumed that the function is of known form with unknown parameters say:

$$M(x) = B_0 + B_1x + B_2x^2 + \ldots$$

where the parameters, B_i, are estimated on the basis of observations $Y_1, Y_2, \ldots, Y_n$ corresponding to observed values $x_1, x_2, \ldots x_n$. The method of least squares, for example, yields the estimators of B_i which minimize the sum of the squared errors.

However cases often arise in which one has little prior knowledge of the actual form of $M(\cdot)$ or one is only interested in trying to estimate the value θ such that

$$M(\theta) = \alpha$$

where α is a specific desired response level. We desire to find a sampling scheme such that $X_n \to \theta$.

Robbins and Monro [Ref. 88] presented the following:

THEOREM: Let $M(x)$ be a given function and α a given constant such that the equation $M(x) = \alpha$ has a uniquely defined root* $x = \theta$. Let $Y(x)$ denote a realization of an experiment at "control level" x. Assume $Y(x)$ has distribution $P(Y(x) \leq Y) = H(Y|x)$ such that $M(x) = \int_{-\infty}^{\infty} Y dH(Y|x)$. (I.e., $M(x) = E(Y|x)$.) Choose X_1 arbitrary and define the recursive relation: $x_{n+1} = x_n + a_n\{\alpha - Y_n(x_n)\}$.

If there exists a positive constant C such that

$$P[|Y(x)| \leq C] = 1$$

and if

$$\text{for some } \delta > 0$$

$$M(x) \leq \alpha - \delta \quad \text{for} \quad x < \theta$$

and

$$M(x) \geq \alpha + \delta \quad \text{for} \quad x > \theta$$

---

* Note that this requires that for some $\delta > 0$, $M(x) \leq \alpha - \delta$ for $x < \theta$ and $M(x) \geq \alpha + \delta$ for $x > \theta$, but does not specifically require that $M(\theta) = \alpha$.

then for $a_n = 1/n$

$$\lim_{n\to\infty} E[(x_n - \theta)^2] = 0 .$$

The procedure of recursively defining x_{n+1} as a function of x_n by

$$x_{n+1} = x_n + a_n(\alpha - Y(x_n))$$

is referred to as the Robbins-Monro method or procedure. (Note that the process is a first order Markov process, although it is in general non-homogeneous.) Papers which followed Robbins and Monro's discussed topics such as convergence, finding the maximum (or minimum) of a function, multidimensional applications, and accelerated processes to name a few.

In the first few years of stochastic approximation survey papers by Derman (1956) [Ref. 25], Schmetterer (1960) [Ref. 101], and Loginov (1966) [Ref. 81] presented major results through their respective date of publication. A text on the subject was attempted by Wasan (1969) [Ref. 129] but received strong criticism because of serious oversights and many misprints.[1] While the aforementioned publications contained only the mathematical formulation, other treatments

---

1 Dupac, V., Book Review, Annals of Mathematical Statistics, v. 41, p. 1131, 1970.

by Fu [Refs. 53, 54] and Wetherill [Ref. 130] contained predominantly practical and intuitive information and little mathematical background. This paper will attempt to present the major results of both mathematical formulation and practical applications and to discuss the intuitive meaning where it is applicable. The list of references is intended to be as complete as possible on the subject. Consequently many of the bibliographical entries are not specifically referenced.

II. MOTIVATING STOCHASTIC APPROXIMATION

In certain applications, as in bioassy, sensitivity testing, or fatigue trials, the statistician is often interested in estimating a given quantile of a distribution function or a level of response. Situations of this type are candidates for solution by stochastic approximation methods. Examples of these situations are:

A. QUANTILE ESTIMATION

Suppose we are testing the resistance of a metallic component to fatigue fracture. Let $F(x)$ denote the probability that a specimen will fail if subjected to x cycles in a trial. Then a specimen, when tested in such a way, represents an observation which takes on a value one or zero depending on whether or not it fractures in x cycles. Thus in the notation of the previous section, $Y(x) = 1$ if the specimen fractures and $Y(x) = 0$ otherwise, so that $M(x) = \int_{-\infty}^{\infty} Y(x)dF(x) = F(x)$. It is of interest to estimate the number of cycles, x, such that for a given α, $F(x) = M(x) = \alpha$.

1. LD_{50}

We wish to administer sample doses of a drug to laboratory animals, say rats, such that we determine the dosage such that 50% of them die on the average. In this case $\alpha = .5$ in our problem formulation and we desire to solve $M(x) = .5$ for x.

B. LEVEL DETERMINATION

Suppose in production it is desired to find the level of some material such that a characteristic, say the viscosity of the finished product, is a pre-determined level. However each batch is subject to impurities and reacts as a stochastic realization. A stochastic approximation scheme may be devised to automatically set and correct the desired input flow to produce the desired results.

C. ROUND-OFF ERRORS

As stated by Schmetterer [Ref. 101] we can consider an application of the RM process for the problem of round-off errors. This problem occurs, for example, if one solves equations by classical iteration process using electronic computers. Define for every real number X a random variable, Y(x), in the following way.

$$P[Y(x) = [x]] = 1 - (x - [x]),$$

$$P[Y(x) = [x] + 1] = x - [x]^{*} \quad .$$

Note that $E[Y(x)] = x$. From here we can deduce as a pattern for more general theorems the following result. If one solves a linear equation by an iterative procedure and modifies it by using for every step of the iteration the

---

* Note that [x] denotes the largest interger contained in x. For example [2.87] = 2.

round-off rule given above, then the modified procedure converges with probability one to a solution of the given equation.

III. THE ROBBINS-MONRO METHOD

In the introduction the first of two theorems from the original Robbins-Monro paper was presented. This first theorem required that the response function, $Y(x)$, be bounded, allowed discontinuities in the function $M(x) = E[Y(x)]$, and did not specifically require that $M(x = \theta) = \alpha$, the desired response level. The second theorem of Robbins-Monro is presented below:

THEOREM [Ref. 88]

Let the sequence a_n be of the form $1/n$ and assume there exists some constant $C > 0$ such that

$$P\{|Y(x)| \leq C\} = 1,$$

and that the conditions

(i.) $M(x)$ is a nondecreasing function,

(ii.) $M(\theta) = \alpha$,

(iii.) $M'(\theta) > 0$

are satisfied. Then defining the recursive relation

$$x_{n=1} = x_n + a_n[\alpha - Y(x)]$$

implies the result

$$\lim_{n\to\infty} E[(x_n - \theta)^2] = 0$$

A. WEAKENING THE CONDITIONS FOR CONVERGENCE

Wolfowitz [Ref. 132] in response to questions of Robbins and Monro showed that if the conditions of the response function, M(x), satisfy

$$|M(x)| \leq C,$$

$$\int_{-\infty}^{\infty} (Y(x) - M(x))^2 dF(Y|x) \leq \sigma^2 < +\infty,$$

along with

$$M(x) < \alpha \quad \text{for } x < \theta,$$

$$M(\theta) = \alpha,$$

$$M(x) > \alpha \quad \text{for } x > \theta,$$

M(x) strictly increasing when $|x - \theta| < \delta$ for some $\delta > 0$ and

$$\inf_{|x-\theta| \geq \delta} |M(x) - \alpha| > 0.$$

Then for x_n defined as in the RM process x_n converges in probability to θ.

B. CONVERGENCE WITH PROBABILITY 1

Kallianpur [Ref. 68] and Blum [Ref. 6], both proved a convergence which is stronger than convergence in mean square (which implies convergence in probability), convergence with probability 1. Blum [Ref. 6] proved that the RM process converges with probability 1 under conditions even weaker than those of Wolfowitz [Ref. 132]. While Wolfowitz required that the regression function $M(\cdot)$ be bounded, Blum only required that it lie between two lines.

BLUMS' THEOREM [Ref. 6]

Let $M(x)$ be the regression function. We assume that $M(x)$ is measurable and satisfies the following conditions:

$$M(x) \leq C + d|x|$$

for some $C, d > 0$

$$\int_{-\infty}^{\infty} (Y - M(x))^2 dF(Y|x) \leq \sigma^2 < +\infty,$$

$$M(x) < \alpha \quad \text{for} \quad x < \theta,$$

and

$$M(x) > \alpha \quad \text{for} \quad x > \theta,$$

$$\inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |M(x) - \alpha| > 0$$

for any pair (δ_1, δ_2), if moreover

$$\sum_{n=1}^{\infty} a_n = +\infty,$$

and

$$\sum_{n=1}^{\infty} a_n^2 < +\infty,$$

then x_n converges to θ with probability 1.

I.e.

$$P\{\lim_{n\to\infty} x_n = \theta\} = 1.$$

C. A FURTHER WEAKENING OF CONDITIONS FOR CONVERGENCE

In 1963 Friedman [Ref. 51] further weakened the requir - ments for convergence with probability 1 by removing the necessity for $M(x)$ and $\sigma^2(x)$ to be bounded by a linear function and a constant respectively. Friedman's theorem is presented here:

THEOREM [Ref. 51]

Let $f(x)$ be a function which is positive and bounded in any finite interval. Let the following conditions be satisfied:

$$M(x) \leq (L\,|x| + K)f(x)$$

for constants $L, K, M > 0$,

$$\sigma^2(x) \leq \sigma^2 f(x),$$

$$M(x) < \alpha \quad \text{for} \quad x < \theta,$$

and

$$M(x) > \alpha \quad \text{for } x > \theta,$$

$$\inf_{\delta_1 \leq |x-\theta| \leq \delta_2} |M(x) - \alpha| > 0$$

for any pair (δ_1, δ_2), then the sequence defined by

$$x_{n+1} = x_n - a_n[\alpha - Y(x_n)]/f(x_n)$$

converges to θ with probability 1.

This theorem of Friedman enables one to construct a convergence process when $|M(x)|$ and $\sigma^2(x)$ are bounded by known functions $f_1(x)$ and $f_2(x)$. One then takes $f(x) = \text{Max}[f_1(x), (f_2(x))^{\frac{1}{2}}]$. This procedure is also applicable where $f(x)$ is decreasing to zero for large values of x. However the convergence is relatively slow.

Later Gladyshev [Ref. 57] simplified the conditions for convergence with probability 1 with the following:

THEOREM

Let M(x) be a measurable function such that

$$\inf_{\varepsilon<|x-\theta|<\varepsilon^{-1}} (x - \theta)\{M(x) - \alpha\} > 0 \quad \text{for all } \varepsilon > 0.$$

Moreover assume there exists a positive number d such that for all x we have

$$E[Y^2(x)] \leq d(1 - x^2)$$

and if $\{a_n\}$ satisfies the conditions previously stated then x_n converges to θ with probability 1 where x_n is defined as in the original RM process.

D. THE MULTIDIMENSIONAL CASE

Blum [Ref. 7] was the first to generalize the Robbins-Monro process to the multidimensional case. He considered the following problem:

Assume that we are given a family of N random variables

$$Y_1(x_1,\ldots x_n), \ \ldots, \ Y_n(x_1,\ldots,x_n)$$

with distribution functions

$$F_1(\cdot),\ldots,F_n(\cdot)$$

moreover assume that

$$M_i(x_1,\ldots x_n) = \int_{-\infty}^{\infty} Y_i dF_i \quad \text{for } i = 1,\ldots,N,$$

(i.e. M_i is the corresponding regression function in the ith dimension).

It is desired to construct a sequence whose limit is the root vector of the system of equations

$$M_i(x_1,\ldots,x_n) = \alpha_i.$$

For simplicity it is assumed that all $\alpha_i = 0$ and that

$$M_i(0) = 0.$$

Let f(x) be a real function that is defined on real N-dimensional space and has continuous first and second derivatives. Let $A(x) = (\partial^2 f/\partial x_i \partial x_j)$ denote the matrix of second derivatives and let $D(x) = (\partial f/\partial x_i)$ denote the vector of the first derivatives.

In matrix form the Robbins-Monro process is of the form $x_{n=1} = x_n + a_n Y_n$ where it is assumed as before that $\sum_{n=1}^{\infty} a_n = \infty$, and $\sum_{n=1}^{\infty} a_n^2 < +\infty$.

Observe the following notation:
Let $U(x) = \langle D(x), M(x)\rangle$ (i.e. the scalar product of D(x) and M(x)).

Let $V_a(x) = E\{\langle Y_x, A(x + Qa\ Y_x)Y_x\rangle\}$ where $0 \leq q \leq 1$.

THEOREM [Ref. 7]

If there exists a real function $f(x)$ with continuous derivatives that satisfies the conditions

$$\text{(i)} \quad f(x) \geq 0,$$

$$\text{(ii)} \quad \underset{||x||>\varepsilon}{\text{Sup}}\ U(x) < 0 \quad \text{for } \varepsilon > 0,$$

$$\text{(iii)} \quad \underset{||x||>\varepsilon}{\text{Inf}}\ |f(x) - f(0)| > 0 \quad \text{for } \varepsilon > 0,$$

$$\text{(iv)} \quad V_a \leq V < +\infty \quad \text{for all } a,$$

then the sequence $\{x_n\}$ as previously defined converges almost surely to zero.

It should also be noted that the multidimensional case is a direct extension of theorems by Derman and Sacks [Ref. 26] and by Gladyshev [Ref. 57] where we think of x_n, Y_n, θ, α, and $M(x)$ as N dimensional vectors and treat multiplication as the vector scalar product. Of interest is a special case when the regression function, $M(x)$ is linear, $M(x) = \alpha$ where M is a symmetric matrix. The following modified RM process was proposed by Dupac [Ref. 33] for this special case.

Here Y is a random vector whose distribution function is $F(Y|Y_n - \alpha)$ where Y_n is a random vector with distribution

function $F(Y_n|x_n)$. Then the following theorem is presented by Dupac.

THEOREM [Ref. 33]

Assume that

$$\int_{R_n} ||Y - M(x)||^2 dF(Y|x) \leq \sigma^2 < +\infty$$

and

$$K_1 = \min_i \lambda_i^2$$

are satisfied where the λ_i are the characteristic roots of the matrix M. Then if $a > 1/2K_1$, then the sequence defined by

$$x_{n+1} = x_n - \frac{a}{n} Y_n$$

converges to θ with probability 1 and

$$E\{||x_n - \theta||^2\} = 0\ (1/n)$$

There are certain situations where the multidimensional case can be reduced to a one dimensional case. The results are by Eppling [Ref. 37] and require general stochastic approximation theorems of the Dvoretzky type.

E. DVORETZKY'S GENERALIZED PROCESS

Dvoretzky [Ref. 36] has suggested that any stochastic approximation procedure may be viewed as an ordinary deterministic (error free) successive approximation method with a noise component superimposed on it at each step. On the basis of this concept a very generalized class of stochastic approximation theorems can be studied.

Assume that $T_n(x_1,\ldots,x_n)$ is a Borel-measurable sequence of transformations from n-dimensional Euclidian space, R_n, into R_1. One may then construct the sequence from the relation

$$x_{n+1} = T_n(x_1,\ldots,x_n) + Z_n$$

where $T_n(x_1,\ldots,x_n)$ is the error free transformation and Z_n is the error. Dvoretzky then proved the following theorem:

THEOREM [Ref. 36]

Let $\{a_n\}$, $\{B_n\}$ and $\{\gamma_n\}$ be sequences of non-negative real numbers, such that,

$$\lim_{n\to\infty} a_n = 0,$$

$$\sum_{n=1}^{\infty} B_n < \infty,$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty.$$

Moreover, assume that the condition

$$|T_n(r_1\ldots,r_n) - \theta| \leq \text{Max } [a_n,(1+B_n)|r_n-\theta|-\gamma_n]$$

is satisfied for all real $r_1,\ldots,r_n$; also that

$$\sum_{n=1}^{\infty} E(Z_n^2) < \infty$$

and

$$E(Z_n|x_1,\ldots,x_n) = 0$$

with probability one.* Then the sequence $\{x_n\}$ defined by

$$x_{n+1} = T_n(x_1,\ldots,x_n) + Z_n$$

converges to the desired quantity, θ, in mean square and with probability 1. I.e.,

$$\lim_{n\to\infty} E[(x_n - \theta)^2] = 0,$$

and

---

* Note that this condition is satisfied if the Z_n are a sequence of independent errors for which $E(Z_n) = 0$ for all n.

$$P \lim_{n\to\infty} x_n = \theta = 1$$

It can easily be shown that the Robbins-Monro procedure is a special case of Dvoretzky's generalized procedure. To do this write the normal R-M relation

$$x_{n+1} = x_n + a_n\{\alpha - Y(x_n)\}$$

as

$$x_{n+1} = x_n + a_n[\alpha - M(x_n) + a_n[M(x_n) - Y(x_n)].$$

Then letting

$$T_n(x_1,\ldots,x_n) = x_n + a_n[\alpha - M(x_n)]$$

and

$$Z_n = a_n[M(x_n) - Y(x_n)]$$

we have the RM procedure in Dvoretzky's format. In a similar manner the Kiefer - Wolfowitz procedure, which will be discussed later, can be written as a special case of Dvoretzky's theorem.

Dvoretzky extended his generalized procedure even further by replacing the sequences a_n, B_n, and γ_n by

non-negative functions $a_n(r_1,\ldots r_n)$, $B_n(r_1,\ldots r_n)$, and $\gamma_n(r_1,\ldots r_n)$ respectively provided that they satisfy the following conditions:

(i) the functions $a_n(r_1,\ldots,r_n)$ are uniformly bounded and $\lim_{n\to\infty} a_n(r_1,\ldots,r_n)=0$ uniformly for all sequences $r_1,\ldots,r_n,\ldots$

(ii) the functions $B_n(r_1,\ldots,r_n)$ are measurable and $\sum_{n=1}^{\infty} B_n(r_1,\ldots,r_n)$ is uniformly bounded and uniformly convergent for all sequences $r_1,\ldots,r_n$.

(iii) the functions $\gamma_n(r_1,\ldots,r_n)$ satisfy $\sum_{n=1}^{\infty} \gamma_n(r_1,\ldots,r_n) = \infty$ uniformly for all sequences $r_1,\ldots,r_n$ for which $\underset{n=1,2,\ldots}{\text{SUP}} |r_n| < L$ where L is a finite number.

The introduction of Dvoretzky's general conditions allowed regression functions of the form $M(x) = -x^2$ or $M(x) = \text{Exp}\ (-x^2)$ to be applicable to stochastic approximation type theorems. The most comprehensive presentation of Dvoretzky stochastic approximation theorems has been by Venter [Ref. 120] in 1966. Venter's theorems generalized the work of Dvoretzky [Ref. 36] and Wolfowitz [Ref. 133] for transforms on the real line, of Derman and Sacks [Ref. 26]

for finite dimensional Euclidian spaces, and of Schmetterer [Ref. 101] for Hilbert spaces.

Block [Ref. 5] had proposed a more general type of stochastic approximation taking place on a normed vector space.

F. EXPECTED SQUARED ERROR

While Blum [Ref. 6], Dvoretzky [Ref. 36], and Dupac [Ref. 32] were establishing conditions under which $b_n = E[(x_n-\theta)^2] \to 0$, others such as Chung [Ref. 14], Hodges and Lehmann [Ref. 64], Kallianpur [Ref. 68], and Schmetterer [Ref. 101] were trying to establish bounds on b_n. (Note that b_n = variance + $(\text{bias})^2$.) Below is Schmetterer's result for the bounded case.

THEOREM [Ref. 101]

Let M(x) be a Borel-Measurable function that satisfies

(i) $P\{|Y(x)| \leq C\}=1$ for some constant $|C|<+\infty$

and

(ii) $(x-\theta)\{M(x)-\alpha\}>0$ for all $x \neq \theta$.

Also there exists an $\varepsilon > 0$, and positive constants C_1, C_2, such that

(iii) $|M(x)-\alpha| \geq C_1|x-\theta|$ for $|x-\theta| \leq \varepsilon$,

(iv) $|M(x)-\alpha| \geq C_2$ for $|x-\theta| > \varepsilon$.

Then

$$b_n \leq b_1 \prod_{i=1}^{n-1}\left(1-\frac{C_3 a_i}{A_{i-1}}\right)+\prod_{i=1}^{n-1}\left(1-\frac{C_3 a_i}{A_{i-1}}\right)\sum_{i=1}^{n-1} a_i^2 e_i\left[\prod_{r=1}^{i}\left(1-\frac{C_3 a_r}{A_{r-1}}\right)\right]^{-1}$$

where $e_i = E\,(y_i-\alpha)^2$, a_0 is defined as 1, $A_n = \sum_{i=1}^{n} a_n$,

and there exists a constant, C_3, such that

$$|M(x)-\alpha| \geq \frac{C_3}{2A_{n-1}}\,|x_n-\theta| \quad \text{with probability i.}$$

As was noted above this theorem holds for the so called "bounded case." This same result holds for the quasilinear case if conditions (i), (iii), and (iv) are replaced by the following conditions:

There exists a C_4 such that

$$E\{[Y(x)-M(x)]^2\} \leq C_4$$

and the quasilinear conditons that there exist C_5 and C_6, $C_5 < C_6$ such that

$$C_5|x_n-\theta| \geq |M(x)-\alpha| \geq C_6|x_n-\theta| \ .$$

Then the above estimate of b_n holds when a_i is substituted for a_i/A_{i-1} and $2C_6$ is substituted for C_3.

Therefore if $a_n=a/n$ where $a > 1/2C_6$ then $b_n = 0(1/n)$. This latter is the most frequently used result.

G. EXPECTED SQUARED ERROR IN THE LINEAR CASE

Hodges and Lehmann [Ref. 64] analyzed in detail the case where it is desired to estimate the value of x for which $M(x)=0$ where it is assumed that $M(x)=\beta x$, and variance $(Y(x)) = \sigma^2$. Then using the RM process to define $\{x_n\}$ yields

$$b_n = E[x_n^2] = x_1^2\left[\prod_{r=1}^{n-1}(1-\beta a_r)\right]^2 + \sigma^2 \sum_{r=1}^{n-1} a_r^2 \prod_{s=r+1}^{n-1}(1-a_s)^2 .$$

In analyzing this expression it becomes obvious that the first term represents the expected bias based on the initial choice, x_1, while the second term represents the variance component of the error variance. Since Chung [Ref. 14] established that under certain conditions the sequence $a_n=c/n$ gives most rapid convergence of x_n to θ, it is of interest to analyze the expression for expected squared error for this family of coefficients.

For the first (expected bias to initial value, x_1) term the expected bias = $O(n^{-2c\beta})$ if $(c/n)^{-1} \geq \beta$ for all n, but becomes quite large if $(c/n)^{-1} < \beta$. Hence as noted by Wetherhill [Ref. 130] it would be more desirable to tend to overestimate $c = \beta^{-1}$ rather than underestimate.

The analysis of the second term is more complicated but Hodges and Lehmann have shown that it is asymptotically equivalent to

$$\frac{\sigma^2 c^2}{n(2c\beta-1)} \qquad \text{if} \quad c > 1/2\beta$$

$$\frac{\sigma^2 \log_e n}{4n\beta^2} \qquad \text{if} \quad c = 1/2\beta$$

and we note that c should not be less than $(2\beta)^{-1}$ because of the large bias which results.

These results give us conditions on the sequence, $a_n = c/n$ in terms of β, the slope of the regression function, such that the bias, resulting from an initial bad guess, rapidly tends to zero with increasing sample size and the expected squared error of x_n is of the order $0(1/n)$. The main shortcoming of the linear model is that we do not know how nearly linear $M(x)$ must be nor how nearly constant variance (x) must be in order that the linear approximation will represent what actually happens. The only evidence on this point consists of a sampling experiment by Teichroew [Ref. 109]. There it was found that the linear theory is in reasonable agreement with the data.

H. RATE OF CONVERGENCE

In a recent paper Komlos and Revesz [Ref. 136] presented estimates of the rate of convergence of the R-M process in a form more concise than any previous result. They considered the case $\alpha = 0$, $\theta = 0$, $a_n = 1/n$ and presented the following estimates.

For the case where there exists $L > 0$ such that

$$P[|Y(x)-M(x)| \leq L] = 1 \quad \text{and} \quad \lim_{x\to\infty} M(x)=M(\infty) > L,$$

if the conditions

$$M(x) < c_1 x + d_1 \qquad \text{if } x > \theta = 0$$

$$M(x) > c_2 x + d_2 \qquad \text{if } x < \theta = 0 ,$$

are satisfied for positive constants c_1, c_2, d_1, d_2,

then $$P[x_n > \varepsilon] \leq e^{-\gamma n}$$

for any $\varepsilon > 0$, where $\gamma = \gamma(\varepsilon) > 0$.

For the case where $M(\infty) < L$ the rate of convergence is much slower than for the previous case. Specifically

$$P[x_n > \varepsilon] \leq \exp\left(-n^{\frac{M(\infty)}{L} - \delta}\right)$$

for any $\delta > 0$, $n > n_0(\delta)$.

I. ASYMPTOTIC NORMALITY

The asymptotic behavior of the higher-order moments an the asymptotic distribution of the random variables defined by the sequence $\{x_n\}$ was first considered in detail by Chung [Ref. 14]. His method is based on the moments of $(x_n-\theta)$ and his results have been widely used in papers on stochastic approximation. Chung's fundamental result is for the "bounded case" and can be stated as follows:

THEOREM [Ref. 14]

Let $M(x)$ be a Borel-measurable function that satisfies the following conditions

$$P\{|Y(x)| \leq C_1\} = 1 \qquad \text{for some constant, } C_1,$$

$$(x-\theta)\{M(x)-\alpha\} > 0 \quad ,$$

$$M(x) = \alpha + \alpha_1(x-\theta) + O(|x-\theta|) \quad ,$$

$$\inf_{|x-\theta|>\delta} |M(x)-\alpha| = K_0(\delta) \quad 0 \quad .$$

and

$$E[(Y(x)-M(x))^2] \leq \sigma^2 < \infty \qquad \text{for all } x \ .$$

Let $a_n = 1/n^{1-\varepsilon}$,

$$\text{where} \qquad \frac{1}{2(1+C_2)} < \varepsilon < \frac{1}{2}$$

and where C_2 is determined by the condition

$$|M(x_n)-\alpha| > C_2 \varepsilon n^{-\varepsilon} |x_n-\theta| \quad .$$

Then for any integer $r \geq 1$

$$\lim_{n\to\infty} n^{(1-\varepsilon)\frac{r}{2}} E[(x_n-\theta)^r] = \begin{cases} 0 & \text{if } r \text{ is odd,} \\ (\sigma^2/2\alpha_1)^{\frac{r}{2}} (r-1) & \text{if } r \text{ is even,} \end{cases}$$

and the random variable $n^{(1-\varepsilon)\frac{1}{2}} (x_n-\theta)$ is asymptotically normally distributed with mean zero and variance $= \sigma^2/2\alpha_1$.

A similar result was obtained by Chung for the quasi-linear case (i.e. $M(x)$ lies between two straight lines with nonvanishing slope). In this case the boundedness assumption is replaced by

$$K|x-\theta| \leq |M(x)-\alpha| \leq K_1|x-\theta| \qquad \text{for } K > 0, \quad K_1 < \infty$$

and

$$E[\{Y(x)-M(x)\}^p] \leq K_2 < \infty \qquad \text{for } p \text{ an even integer.}$$

Then using $a_n = c/n$ where $c > 1/2K$, the distribution of $n^{1/2}(x_n-\theta)$ tends to normal with mean zero and variance = $\sigma^2c^2/(2\alpha_1 c-1)$.

For the special case where M(x) is linear Chung proves asymptotic normality by using characteristic functions in a very concise proof. While Hodges and Lehmann [Ref. 64] improved some of Chung's results, Sacks [Ref. 95] utilized a central limit theorem for dependent random variables to obtain more general and more complete results about the asymptotic normality of X_n. Below is a theorem of Gladyshev [Ref. 57] that is a strengthened form of Sack's fundamental result.

THEOREM [Ref. 57]

Assume that the following conditions are satisfied:

$$\inf_{\varepsilon<|x-\theta|<1/\varepsilon} (x-\theta)(M(x)-\alpha) > 0 \qquad \text{for } \varepsilon > 0 ,$$

$$M(x) = \alpha + \alpha_1(x-\theta) + \delta(x,\theta)(x-\theta) ,$$

there exists a $d > 0$, such that for all x,

$$E[Y^2(x)] \leq d(1+x^2) ,$$

$$\lim_{n\to\infty} E[(Y(x)-M(x))^2] = \rho > 0 ,$$

$$\lim_{N\to\infty} \lim_{\varepsilon\to 0} \sup_{|x-\theta|<\varepsilon} E[(Y(x)-M(x))^2\Phi_N(x)] = 0 ,$$

where

$$\Phi_N(x) = \begin{cases} 1 & \text{for } |Y(x)| > N , \\ 0 & \text{for } |Y(x)| \leq N , \end{cases}$$

and

$$a_n = An^{-1} \text{ is such that } A\alpha_1 > 1/2 .$$

Then the distribution of $n^{1/2}(x_n-\theta)$ tends to normal with mean zero and variance = $A^2(2A\alpha_1-1)^{-1}\pi\rho$.

J. SELECTION OF STEP SIZE, a_n

As we have noted thus far the sequence $\{a_n\}$ must essentially have the same asymptotic behavior as the harmonic series, $1/n$, which satisfies the conditions

$$\sum_{n=1}^{\infty} a_n = \infty \quad , \quad \text{and} \quad \sum_{n=1}^{\infty} a_n^2 < \infty .$$

We can intuitively see that the first condition is necessary to guarantee that the sequence, $\{x_n\}$, does not get trapped in any finite interval while the second condition is necessary for the convergence of the expected squared error term. However it is reasonable to ask if there is a sequence, $\{a_n\}$, which minimizes $E[(x_n-\theta)^2]$ after some fixed number of observations, say N. Dvoretzky [Ref. 36] solved this problem for the Robbins-Monro Process.

THEOREM [Ref. 36]

Assume that a random variable, Y(x), satisfies the conditions

(i) $E[Y^2(x)] \leq \sigma^2 < \infty$

and assume that $M(x)$ is such that

$$\text{(ii)} \quad 0 < A \leq \frac{M(x) - \theta}{x - \theta} \leq B < \infty$$

and if it is known that

$$\text{(iii)} \quad |x_1 - \theta| \leq C \leq \sqrt{\frac{2\sigma^2}{A(B - A)}} .$$

Then if the sequence, $a_n = \frac{Ac^2}{\sigma^2 + nA^2}$ is used in the Robbins-Monro process, then the resultant

$$E[(x_n - \theta)^2] \leq \frac{\sigma^2 c^2}{\sigma^2 + (n-1)A^2 c^2}$$

is obtained. The choice of $\{a_n\}$ here is optimal in the minimax sense in that for any other choice of $\{a_n\}$ there exist $Y(x)$ and x_1 that satisfy conditions (i) and (iii) for which the above bound on expected squarred error does not hold.

Now it is obvious that this information is of limited use to the experimenter who has little a priori information with which to choose a_n. Therefore for practical choice of the sequence, a_n, the reader is directed to Section V.A. where this problem is discussed.

K. ACCELERATING CONVERGENCE

When the initial guess, x_1, is far from the desired value of θ, the Robbins-Monro procedure approaches θ very

slowly because we are taking smaller and smaller steps. Kesten [Ref. 69] proposed the method of accelerating the convergence of a stochastic approximation algorithm based on not decreasing the step size, a_n, if the difference $(x_n - x_{n-1})$ has the same sign as $(x_{n-1} - x_{n-2})$, and decreasing the step size if the signs differed, indicating that we may be in the region of θ. (Higher order schemes are also proposed.)

It was shown that there exists a θ', not necessarily identical with θ, about which fluctuations in sign occur more frequently in a finite number of trials. The value of $x = \theta'$ is defined by the intersection of the line $Y(x) = \alpha$ and the locus of medians of the densities $\frac{d}{dy}(F(Y|x))$ for any x. If the density $\frac{d}{dy}(F(Y|x))$ is symmetric, then $\theta' = \theta$. Even if the fluctuations occur about a θ', different from θ, x_n still converges in probability to θ as Kesten proved.

Authors such as Odell [Ref. 87], Sinha and Griscik [Ref. 105], Sielken [Ref. 104], and Newbold [Ref. 86] have presented accelerated stochastic approximation methods of their own and have compared them with the original R-M method and Kesten's method.

Another method of accelerating convergence was proposed by Fabian [Ref. 40]. This method is an analog of the method of steepest ascent (descent). Fabian proposed that the step a_n be determined in the following manner; for given x_n and y_n one makes a series of observations, V_j, (where the

observations are assumed to be independent of x_n and y_n) of the quantity $M(x_n + jay_n)$ for $j = 1,2...$ until sign $V_1 = ... = \text{sign } V_{j-1} = \text{sign } V_j = -\text{sign } V_{j+1}$. Then choose $a_n = ja$. (Note here $\alpha = 0 = M(\theta)$.) Fabian proved that under very general conditions on V_j iteration methods converge with probability 1.

Authors who are interested in the practical or experimental aspects of stochastic approximation have suggested that the approximation method be carried out in two stages. The first stage would take large steps to estimate the region of interest while the second stage would take progressively smaller steps and represents the fine tuning stage. (See Davis [Ref. 22], Wetherill [Ref. 130], and Goodman, Lewis and Robbins [Ref. 58].)

L. CONFIDENCE INTERVALS AND STOPPING TIMES

After k iterations it may be desired to obtain an estimate of γ and d such that

$$P(|x_{k+1} - \theta| \leq d) \geq 1 - 2\gamma .$$

Farrell [Ref. 50] did some of the first work on confidence intervals of bounded length but required a priori knowledge of a bounded interval containing θ.

The subject of stopping times of a non-parametric nature is an almost untouched area. Farrell stated that Mrs. Nancy Tapper, Cornell University, had been studying closed stopping rules and bounded length confidence interval procedures for

the median of a distribution function. However very little has appeared in stochastic approximation literature concerning stopping rules.

The most general discussion of stopping times based on the asymptotically normal result was recently presented by Sielken [Ref. 104] and is stated below.

Using the definition,

$$Z(x) = Y(x) - M(x)$$

consider the following conditions;

(1) γ is a positive constant less than 1/2.

(2) The sequence $\{c_n\}$ of positive constants is such that $c_n\gamma^n \to c$ as $n \to \infty$, for some $0 < c < \infty$.

(3) The sequence $\{a_n\}$ has the form An^{-1} where A is a constant such that $2A\alpha_1 > 1$.

(4) M is a Borel-measurable function.

(5) For $\varepsilon > 0$, $\inf_{\varepsilon<x-\theta\varepsilon^{-1}} M(x) - \alpha > 0$ and $\sup_{\varepsilon<\theta-x<\varepsilon^{-1}} M(x) - \alpha < 0$.

(6) For some constants K_1 and K_2 $|M(x) - \alpha| \leq K_1 + K_2|x-\theta|$ for all x.

(7) $\sup_x E[|Z(x)|^2] = W$.

(8) $\lim_{x\to\theta} E[|Z(x)|^2] = E[Z(\theta)^2] = \sigma^2 > 0$.

(9) $\lim_{R\to\infty} \lim_{\varepsilon\to 0^+} \sup_{|x-\theta|<\varepsilon} \int_{|Z(x)|>R} |Z(x)|^2 dP = 0.$

(10) For some positive constants g and α_1,

if $|x-\theta| < g$,

then $M(x) = \alpha + \alpha_1(x-\theta) + \delta(x)$,

where $\delta(x) = o(|x-\theta|)$ as $|x-\theta| \to 0$.

(11) The distribution function of Y(x), denoted $F(Y|x)$, is such that for every y, $F(y|\cdot)$ is Borel-measurable.

and

(12) There exists $\varepsilon > 0$ such that for every positive integer r

$$\sup_{|x-\theta|<\varepsilon} E[|Z(x)|^r] < \infty.$$

Then assuming that a $100(1 - 2\gamma)\%$ confidence interval on θ of length 2d is desired, the proposed stopping time for the R-M process is denoted $N_{d,\gamma,1}$ where $N_{d,\gamma,1}$ is the smallest positive integer, n, such that

$$n \geq K_\gamma^2 A^2 S_{n,1}^2/(2At_{n,1} - 1)d^2.$$

The principle results of Sielken are:

THEOREM [Ref. 104]

If conditions (1) - (12) above are satisfied then

$$\lim_{d\to 0} N_{d,\gamma,1}/[K^2 A^2\sigma^2/(2A\alpha_1 - 1)d^2] = 1 \quad ,$$

with probability 1,

and

$$\lim_{d \to 0} P(|X_{N_{d,\gamma,l}+1} - \theta| \leq d) = 1 - 2\gamma.$$

Sielken has stated that the limit in the theorem can be interpreted as either:

a. The level of the sequentially determined bounded length confidence interval converges to the prescribed level, $1-2\gamma$, as the desired length, $2d$, converges to zero; or

b. The probability that the error in the final estimate of θ is less than or equal to d converges to the prescribed probability, $1 - 2\gamma$, as $d \to 0$.

M. DYNAMIC STOCHASTIC APPROXIMATION

Fabian [Ref. 39] and Dupac [Ref. 34] have considered the case where the desired level, θ, changes during the iteration process. The following discussion is by Fu [Ref. 53] based on Dupac's presentation.

Let $M_n(x) = M(x - \theta_n + \theta_1)$ such that θ_n is the unique root of $M_n(x) = 0$. Let $\{a_n\}$ be a sequence of positive numbers, and let x_1 be an arbitrary random variable.

Define: $x_{n+1} = x_n' - a_n Y(x_n')$,

where

$$x_n' = (1 + n^{-1})x_n,$$

$$E[Y(x_n')|x_1,\ldots x_n] = M_{n+1}(x_n'),$$

and

$$\text{var}[Y(x_n')|x_1,\ldots,x_n] = \sigma^2 < +\infty.$$

The meaning of the above algorithm for computing x_{m+1} with the modified x_n, i.e. x_n', is that when we get an estimate, x_n, of θ_n, we make a correction for trend to obtain x_n' before computing x_{n+1}. It will be seen by the following theorem that the use of this modified algorithm is justified when θ_n is a linear (or nearly linear) function of n.

THEOREM [Ref. 34]

Assume that the following conditions are satisfied:

(i) $M(x) < 0$ for $x < \theta_1$ and $M(x) > 0$ for $x > \theta_1$.

(ii) There exist K_0, K_1 such that

$$K_0|x - \theta_1| \leq |M(x)| \leq K_1|x - \theta_1| \quad \text{for all } x.$$

(iii) $a_n = a/n^{\alpha}$, for $a > 0$, $\frac{1}{2} < \alpha < 1$.

(iv) θ_n varies in such a way that

$$\theta_{n+1} - (1 + n^{-1})\theta_n = O(n^{-\omega}) \quad \text{for } \omega > \alpha$$

and

(v) $E(x_1^2) < +\infty$.

Then $(x_n - \theta_n)$ approaches zero in the mean and

$$E[(x_n - \theta_n)^2] = \begin{cases} O(n^{-\alpha}) & \text{for } \tfrac{1}{2} < \alpha \leq 2/3 \\ O(n^{-2\omega^{-\alpha}}) & \text{for } 2/3 < \alpha < 1. \end{cases}$$

The mean square convergence, as well as convergence with probability 1, can be deduced from Dvoretzky's theorem, even under slightly more general conditions on θ_n. A similar modification to the Kiefer-Wolfowitz procedure is indicated to solve for a moving maximum of a regression function.

An interesting algorithm is presented in Fu's book [Ref. 53] for learning of slowly time varying parameters using dynamic stochastic approximation. Here Kesten's accelerated scheme [Ref. 69] is coupled with Dupac's dynamic process to improve the rate of convergence.

N. CONTINUOUS STOCHASTIC APPROXIMATION

In order to obtain a continuous version of the stochastic approximation method, one can replace the difference recursive relation in the discrete case with a stochastic differential equation. Again letting the desired level of response, α, be equal to zero, one obtains the general expression

$$\frac{d}{dt} X(t) = -a(t)Y(t,X(t)),$$

where a(t) satisfies the conditions

$$\int_0^\infty a(t)dt = \infty \quad \text{and} \quad \int_0^\infty a^2(t)dt < \infty.$$

The above relation determines a continuous process for stochastic approximation of the solution to the equation $M(x) = 0$. Driml and Nedoma [Ref. 31] proved that the process converges when $Y(t,x)$ is monotonic in x and when $Y(t,x)$ is of the form $Y(t,x) = M(x) + h(t)$ where $h(t)$ is an ergodic process with zero mean. In both cases the function $X(t)$ approaches the desired value, 0, with probability 1 as $t \to \infty$. In the proof by Driml and Nedoma

$$a(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq 1 \\ 1/t & \text{for } t > 1. \end{cases}$$

0. EXTENSIONS OF CONTINUOUS STOCHASTIC APPROXIMATION

As was experienced in the discrete case the one dimensional continuous case can be extended to the multidimensional case. However many theorems which are valid for the one dimensional case are not valid for the multidimensional case which depends heavily on stationary point theorems. (I.e. theorems concerning a point x_0 of some space X for which $F(x_0) = x_0$ where F maps X into X.) For a discussion of these theorems see Driml and Hans [Ref. 30] and Hans and Spacek [Ref. 61].

One representation using continuous stochastic approximation is by Kitagawa [Ref. 71] who formulated a Robbins-Monro

model where the Brownian motion process is used to represent the random disturbances inherent in the observations.

IV. FINDING THE MAXIMUM OF AN UNKNOWN REGRESSION FUNCTION: THE KIEFER-WOLFOWITZ METHOD

A problem of practical importance with a regression function, $Y(x)$, is to estimate the value of x, say θ, at which the expectation of $Y(x)$, denoted $M(x)$, is a maximum. To intuitively introduce the method consider the following argument from Wetherill [Ref. 131].

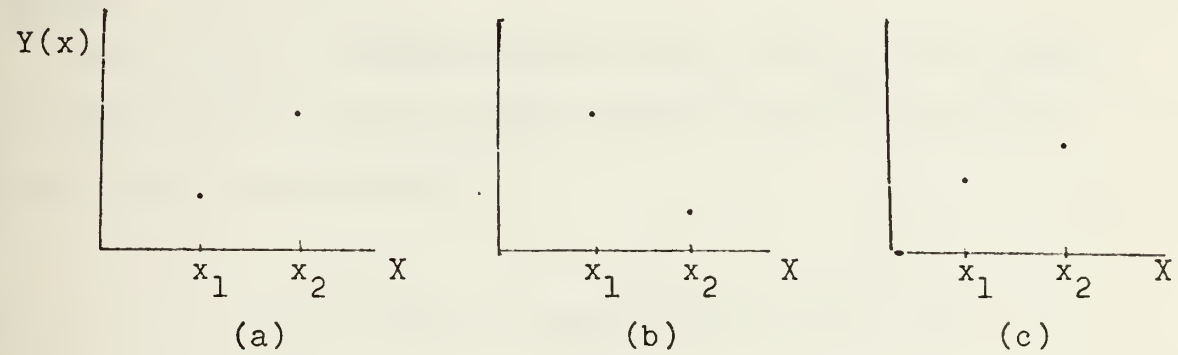


Figure 1

Suppose two observations, $y(x_1)$ and $y(x_2)$, are taken a values x_1 and x_2 where $x_1 < x_2$. Then

(a) If $y(x_1) < y(x_2)$ one expects the maximum level, θ, to be at a value $\geq x_2$.

(b) If $y(x_1) > y(x_2)$ one expects the maximum level, θ, to be at a value $\leq x_1$.

(c) If $y(x_1)$ is about equal to $y(x_2)$ more observations are necessary to determine the region of interest.

Thus it would be reasonable to take further observations in the direction indicated by the slope of the two Y values and the distance moved along the x-axis, before taking further observations, should be proportional to the difference between $y(x_1)$ and $y(x_2)$. Using this basic idea and the initial results of Robbins and Monro, Kiefer and Wolfowitz [Ref. 70] defined the following procedure for stochastic approximation of the maximum of a regression function.

THEOREM [Ref. 131]

Let $M(x)$ be a regression function and $F(Y|x)$ a family of distribution functions and assume that the following conditions are satisfied:

$$\int_{-\infty}^{\infty} (Y(x) - M(x))^2 dF(Y|x) \leq \sigma^2 < +\infty$$

and assume that $M(x)$ is strictly increasing for $x < \theta$, and that $M(x)$ is strictly decreasing for $x > \theta$.

Let $\{a_n\}$ and $\{c_n\}$ be infinite sequences of positive real numbers such that

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n c_n < \infty, \quad \sum_{n=1}^{\infty} a_n^2 c_n^{-2} < \infty$$

(for example: $a_n = n^{-1}$ and $c_n = n^{-1/3}$). Then the recursive scheme defined by

$$x_{n+1} = x_n + \frac{a_n}{c_n} [Y(x_n + c_n) - Y(x_n - c_n)]$$

converges in probability to the maximum, θ, of the regression function $Y(x)$ if three regularity conditions are satisfied. They are listed here with their intuitive meanings.

Condition 1.* There exist positive β and B such that $|x_1 - \theta| + |x_2 - \theta| < \beta$ implies $|M(x_1)-M(x_2)| < B|x_1-x_2|$ for all x_1, x_2. This says if the function, $M(x)$, has a derivative, it must be zero when $x = \theta$; as a result the derivative must be bounded in the neighborhood of θ.

Condition 2. There exist positive ρ and R such that $|x_1 - x_2| < \rho$ implies $|M(x_1) - M(x_2)| < R$. In other words if $M(x)$ increases too abruptly in certain regions, there exists a positive probability that it may reach $+\infty$ or $-\infty$; as a result, the Lipschitz condition must be satisfied.

Condition 3. For every $\delta > 0$, there exists a positive $\pi(\delta)$ such that $|x - \theta| > \delta$ implies $\inf_{\frac{1}{2}\delta>\varepsilon>0} \frac{M(x + \varepsilon) - M(x - \varepsilon)}{\varepsilon} > \pi(\delta)$. Thus if $M(x)$ is a very flat function the rate of motion toward θ is small. As a result, the absolute value of the derivative must be bounded below.

---

* As Blum later proved [Ref. 8], the above theorem holds even when Condition 1 is not satisfied.

While these regularity conditions seem restrictive it is only necessary that they hold in an interval $[c_1, c_2]$ where it is known a priori that $c_1 \leq \theta \leq c_2$. Suppose, however that some proposed level, $a_n \pm c_n$, lies outside the interval, $[c_1, c_2]$ and one cannot take an observation at that level. If one then moves x_n so that the offending $x_n \pm c_n$ is at the boundary (c_1 or c_2) we may proceed as directed and the conclusions remain valid.

A. CONSTANT COEFFICIENTS

Burkholder [Ref. 12] proved that under certain conditions, the Kiefer-Wolfowitz procedure can still be used if c_n is held constant for all n at a particular value, c_0. X_n is then asymptotically normally distributed with variance proportional to n^{-1}. This result is difficult to use in practice since there will rarely be enough information about the response curve to choose c_0 as required by Burkholder.

B. CONVERGENCE WITH PROBABILITY 1

The Kiefer-Wolfowitz process is a special case of the Dvoretzky process. (I.e. the process can be written as the sum of a deterministic term and an error term.) This can be seen by writing

$$x_{n+1} = x_n + \frac{a_n}{c_n} [M(x_n + c_n) - M(x_n - c_n)] + z_n,$$

where the error term is

$$z_n = \frac{a_n}{c_n}[Y(x_n + c_n) - M(x_n + c_n) - Y(x_n - c_n) + M(x_n - c_n)].$$

It follows from a theorem by Dvoretzky that the Kiefer-Wolfowitz procedure converges with probability 1 and in mean square under conditions weaker than those imposed by Kiefer and Wolfowitz. Burkholder [Ref. 12] also proved convergence with probability 1 using a somewhat different approach. Later Venter [Ref. 122] showed that the K-W method converges almost surely to the maximum if this is the only stationary point of the surface and some other conditions are satisfied. This result is stronger, in a sense, than those existing previously.

C. MULTIDIMENSIONAL KIEFER-WOLFOWITZ

Let $(X_1,\ldots,X_N)$ be a family of random variables; let $F_{x_1,\ldots,x_N}$ be the corresponding distribution function; and let $M(x_1,\ldots,x_N)$ be the corresponding regression function. We then desire to find a vector $X = \theta$, for which the regression function is maximal. Assume that $M(x)$ has a unique maximum at the point $x = \theta$.

Blum [Ref. 7] constructed a multidimensional K-W process in the following manner. Let $X \in R_N$ and let $(e_1,\ldots,e_N)$ be an orthonormal basis in R_N. Then for some real $c > 0$, we make $N + 1$ observations of the random variable $Y(\cdot)$,

$$Y(x),\ Y(x + ce_1),\ Y(x + ce_2),\ldots,\ Y(x + ce_N)$$

and consider the vector

$$Y_{x,c} = [\{Y(x = ce_1) - Y(x)\},\ldots,\{Y(x + ce_N)-Y(x)\}].$$

Then beginning with some arbitrary vector, x_1, construct the sequence

$$x_{n+1} = x_n + \frac{a_n}{c_n} Y(x_n),$$

where

$$Y(x_n) \text{ denotes } Y_{x_n,c_n} .$$

Denote the vector of first derivatives of M(x) by D(x), and the matrix of second derivatives by A(x). Then the following theorem by Blum is presented;

THEOREM [Ref. 7]

Let $\{a_n\}$ and $\{c_n\}$ be sequences of positive real numbers that satisfy:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n c_n < \infty, \quad \text{and} \quad \sum_{n=1}^{\infty} a_n^2 c_n^{-2}.$$

Moreover assume that Y(x) and M(x) are such that

$$M(Y(x)^2) \leq \sigma^2 < \infty,$$

$M(\cdot)$ is continuous together with its first and second derivatives, and for any $\varepsilon > 0$, there exists a $\rho(\varepsilon) > 0$ such that

$$||x|| > \varepsilon \quad \text{implies that}$$

$$M(x) \leq -\rho(\varepsilon), \quad \text{and}$$

$$||D(x)|| \geq \rho(\varepsilon),$$

where the partial derivatives $\partial^2 M(x)/\partial x_i \partial x_j$ are bounded for all $i, j = 1,\ldots,N$.

Then the sequence $\{x_n\}$ as previously defined converges to $\theta = 0$ with probability 1. Note that each step in Blum's algorithm requires $N + 1$ observations. Gray [Ref. 59] proved that the multidimensional K-W process defined by

$$x_{n+1} = x_n + \frac{a_n}{c_n}\left[Y^+_{x_n,c_n} - Y^-_{x_n,c_n}\right]$$

also converges with probability one where

$$Y^+_{x_n,c_n} = \{Y(x_n + c_n e_1),\ldots,Y(x_n + c_n e_N)\},$$

$$Y^-_{x_n,c_n} = \{Y(x_n - c_n e_1),\ldots,Y(x_n - c_n e_N)\},$$

which requires $2N$ observations in each step.

D. ASYMPTOTIC PROPERTIES OF K-W PROCESS

The first results concerning asymptotic properties of the Kiefer-Wolfowitz process were obtained by Derman [Ref. 24] and Dupac [Ref. 32] based on the lemmas of Chung [Ref. 14]. Sacks [Ref. 95] has discussed conditions for asymptotic normality of x_n. If c_n is chosen to tend to zero, then the asymptotic variance of x_n can never be made as small, in order of magnitude, as Burkholder's result of being proportional to n^{-1} with $c = c_0$ a constant. The most general results without a priori assumptions about the length of the interval containing the point $x = \theta$ have been obtained by Sacks.

THEOREM [Ref. 95]

Let $M(x)$ be a measurable function with a unique maximum at $x = \theta$, and assume that this function satisfies the conditions:

(i) $$\inf_{\substack{\varepsilon_1 \le |x-\theta| \le \varepsilon_2 \\ 0<\varepsilon<\varepsilon_0}} \frac{(x-\theta)(M(x-\varepsilon) - M(x+\varepsilon))}{\varepsilon} > 0,$$

where $0 < \varepsilon_0 < \varepsilon_1 < \varepsilon_2 < \infty$;

(ii) for all x, $M(x) = \alpha_0 - \alpha_1(x - \theta)^2 + \delta(x,\theta)$, where $\alpha_1 > 0$ and $\delta(x,\theta) = o(|x-\theta|^2)$ as $|x-\theta| \to 0$;

(iii) for some $c_0 > 0$, there exists positive constants K_1 and K_2, such that for all x and all c for which $0 < c \le c_0$

$$K_1(x-\theta)^2 \leq (x-\theta)[M(x-c)-M(x+c)]c^{-1} \leq K_2(x-\theta)^2;$$

(iv) For every $\varepsilon > 0$ there exists a $c_\varepsilon > 0$ such that for all c satisfying $0 < c \leq c_\varepsilon$ and all x satisfying $|x-\theta| < c$

$$|\delta(x-c,\theta) - \delta(x+c,\theta)|c^{-1} \leq \varepsilon|x-\theta|.$$

Further assume

$$\lim_{x\to\theta} E[\{Y(x) - M(x)\}^2] = \sigma^2/2$$

and

$$\lim_{R\to\infty} \lim_{\varepsilon\to 0} \sup_{|x-\theta|<\varepsilon} \int_{|Y(x)-M(x)|>R} (Y(x)-M(x))^2 dP = 0$$

Then if $a_n = An^{-1}$, where $A > 1/2K_1$, the random variable $n^{\frac{1}{2}}c_n(x_n - \theta)$ is asymptotically normally distributed with

$$\text{Mean} = 0$$

$$\text{Variance} = \sigma^2A^2(8\alpha A - 1)^{-1}$$

Sacks, in the same paper, also gave the similar asymptotic limiting distribution for the multidimensional K-W process.

E. MAXIMUM SAMPLE EXCURSIONS IN KIEFER-WOLFOWITZ PROCESS

When we seek a maximum or minimum using the Kiefer-Wolfowitz process the possibility arises that we may be working with a function with more than one local maximum or

that we do not want to reduce the performance, M(x), below some minimum level. The value of x corresponding to this level may not be known. In both of these cases we may wish to limit the excursions to some given multiple or function of $|x_1 - \theta|$, with a high probability, while still being certain that $x_n \to \theta$ with probability 1. To accommodate this situation Kushner [Ref. 79] presented estimates of the following form:

For any $m < \infty$ and even integer r,

$$P\left[\max_{m \geq n \geq N} |x_n - \theta| > \varepsilon\right] < \left[E(x_n - \theta)^r + \delta_{N_r}\right]/\varepsilon^r,$$

where δ_{N_r} depends on the sequences a_n and c_n and can be made arbitrarily small for each fixed N and r, while $x \to \theta$ with probability 1 is still ensured.

F. ACCELERATED CONVERGENCE FOR THE K-W PROCESS

As in the case of the Robbins-Monro process, the rate of convergence of the K-W process can be increased by using Kesten's algorithm [Ref. 69] (See Sec. III.J). Another method for accelerating convergence was proposed by Fabian [Ref. 40] who later showed [Ref. 45] that the multidimensional K-W procedure for functions, f, sufficiently smooth at θ, the point of minimum (or maximum) can be modified in such a way as to be almost as speedy as the R-M method. This modification consists of making more observations at every step and of utilizing these to eliminate the effect of all

derivatives $\partial^j f/\partial x_i^j$, $j = 3,5,7,\ldots,s-1$. Let δ_n be the distance from the approximated θ after n observations. Under similar conditions on f as those used by Dupac [Ref. 32] the result $E\delta_n^2 = O(n^{-s/(s+1)})$ can be obtained. Under weaker conditions it was proved that $\delta_n^2 n^{s/(s+1)-\varepsilon} \to 0$ with probability 1 for every $\varepsilon > 0$.

In a follow-up paper Fabian [Ref. 46] noted that there are many designs, d, which achieve the speed of $E\delta_n^2$ as stated above. He derived the dependence relation of d on

$$\text{Lim } n^{s/(s+1)} E\delta_n^2$$

so that one may choose the design which minimizes $E\delta_n^2$.

In yet a third paper in this series by Fabian [Ref. 48], the results of a design which minimizes $E\delta_n^2$ is utilized and Fabian achieved the result

$$E||x_n - \theta||^2 = o(t_n^{-1}\log^3 t_n)$$

where t_n equals the number of observations necessary to construct $x_1, x_2, \ldots, x_n$.

G. THE CONTINUOUS KIEFER-WOLFOWITZ PROCESS

As with the Robbins-Monro method we have a continuous analog of the Kiefer-Wolfowitz method. Let us consider a method, as discussed in Loginov's survey [Ref. 81], for an ergodic random process Y_t. Let x denote an N-dimensional

vector with coordinates $x_1,\ldots,x_N$ in N-dimensional Euclidian space with orthonormal basis $e_1,\ldots,e_N$. Then the regression function is $M(x) = E[Y_t(x)]$. Moreover assume that

$$y_{i,t}[x,c(t)] = y_t[x + c(t)e_i] - y_t[x-c(t)e_i],$$

where $c(t)$ is some positive function. Then the continuous K-W method of determining a minimum point for a regression function is described by the equation

$$\frac{dx_{i,t}}{dt} = -a(t)I_{i,t}c^{-1}(t)y_{i,t}[x_t,c(t)]$$

with initial conditions $x_{i,0} = x_i(0)$, for $i = 1,2,\ldots,N$, where

$$I_{i,t} = 1 - G_i^{+}(x_{i,t})(F_i^{+}) - G_i^{-}(x_{i,t})F_i^{-}(y_{i,t}).$$

Here G_i^{+} is a monotonic function with derivative bounded on $[b_i - \delta, b_i]$ and

$$G_i^{+}(x) = \begin{cases} 0 & \text{for} \quad x \leq b_i - \delta, \\ 1 & \text{for} \quad x = b_i, \end{cases}$$

and G_i^{-} is a monotonic function with derivative bounded on $[a_i, a_i + \delta]$ and

$$G_i^{-}(x) = \begin{cases} 0 & \text{for} \quad x \geq a_i + \delta, \\ 1 & \text{for} \quad x = a_i, \end{cases}$$

and $F_i^+(y) = 1 - y/\varepsilon c(t)$ and $F_i^-(y) = 1 + y/\varepsilon c(t)$ for $\varepsilon > 0$.

The essential difference between the original discrete Kiefer-Wolfowitz method and this continuous version is the fact that here the observations need not be independent as they were in the discrete case. The term $I_{i,t}$ serves to limit the variable X_i to the interval $[a_i, b_i]$.

Sakrison proved the following convergence theorem for the continuous K-W process.

THEOREM [Ref. 92]

Represent y_t in the form

$$y_t(x) = \sum_{j=1}^{N} g_j(x) V_{j,t}$$

where $V_{j,t}$ are ergodic random processes that are bounded with probability one, while $g_j(\cdot)$ are functions whose second partial derivatives with respect to x_i are bounded.

Now let D_{t+p} denote any of the random processes $V_{e,t+p}$ or $V_{e,t+p}V_{m,t+p}$ $(e, m = 1,2,\ldots,N)$. Moreover let F_t be any bounded functional defined on the processes $V_{e,\tau}$ $(\tau \leq t)$ and $B_{FD}(\rho) = M\{(F_t - M(F_t))(D_{t+p} - M(D_{t+p}))\}$ be such that

$$|B_{FD}(\rho)| \leq \sigma_F \sigma_D (K_1/\rho^2)$$

where $K_1 < +\infty$. Assume that the regression function, $M(x)$ satisfies the conditions

$$(\text{grad } M(z)|_{z=x}, x-\theta) \geq K_2 \, ||x-\theta||^2 ,$$

$$||\text{grad } M(x)||^2 \leq K_3 \, ||x-\theta||^2$$

for $0 < K_2 < K_3 < +\infty$,

$$\left|\frac{\partial^3 M}{\partial x_i^3}\right| \leq P$$

for $i = 1,2,\ldots,N$. Then if the relations

$$\int_0^\infty a(t)dt = \infty$$

$$\int_0^\infty a(t)\, c^2(t)dt < \infty$$

$$\int_0^\infty a(t)\, a(1/2t)dt < \infty$$

$$\int_0^\infty a(t)\, c^{-1}(t)dt < \infty$$

hold for the functions $a(t)$ and $c(t)$, the solution of the stochastic differential equation converges to θ in mean square, i.e.

$$\lim_{t\to\infty} E\{||x_t - \theta||^2\} = 0.$$

An example of functions satisfying the above conditions are

$$a(t) = \frac{a}{(t+1)^{\alpha}} \qquad \text{and} \qquad c(t) = \frac{c}{(t+1)^{\gamma}} ,$$

where

$$\tfrac{1}{2} < \alpha \leq 1 \qquad \text{and} \qquad \gamma > \tfrac{1}{2}(1 - \alpha).$$

Example (by Sakrison [Ref. 92]).

If $\alpha = 1$ and $\gamma \geq \tfrac{1}{4}$,

then

$$E\{||x_t - \theta||^2\} = O(1/n^{\frac{1}{2}}).$$

It is not difficult to see that in the continuous case the requirements of the theorems are considerably more stringent than those in the discrete case. Here constrain s are imposed on the process itself, not just on the regression function. This is the fundamental difference betweer discrete and continuous stochastic approximation methods.

V. PRACTICAL ASPECTS

A. CHOICE OF a_n

In Section III.J. a theorem of Dvoretzky [Ref. 36] was presented giving a formulation for the sequence, a_n, which is optimal in the minimax sense. However this formulation contains parameters which will in general be unknown to the experimenter. The need then arises for a method of choosing a sequence.

Hodges and Lehmann [Ref. 64] recommended using coefficients of the form $a_n = c/n$ where c is chosen to minimize the asymptotic variance, $\sigma^2c^2/(2\alpha_1 c - 1)$. This leads to choosing $c = 1/\alpha_1$ where α_1 is the slope of the response function, M(x), at the desired level of $x = \theta$. (I.e. choose $c = 1/\alpha_1$, where $\alpha_1 = M'(\theta)$.) This does not reduce the experimenter's dilemma since it requires a priori estimation of another unknown parameter. It does however provide a basis for sensitivity analysis on expected squarred error based on changes in the multiplier, c, in terms of α_1. Computer simulations were performed by Hodges and Lehmann [Ref. 64] and by Wetherill [Ref. 130] with very similar results.

In general choosing $c \leq 1/2\alpha_1$ should be avoided since the asymptotic behavior is unknown and simulation experiments indicate that large biases exist when c is chosen to be too small. Similarly when c is chosen too large the asymptotic

variance increases, however it increases slowly for $ca_1 > 1$. Thus when the value of a_1 is unknown it would be more desirable to overestimate the value of c than to underestimate c.

In the special case where M(x) is linear it is easily shown that $a_n = c/n$, with $c = 1/M'(\theta)$, is a desirable choice. Consider the case $M(x) = bx$ where it is desired to sequentially arrive at the value of x where $M(x) = 0$. Without loss of generality let $\theta = 0$. Thus the value of $x = \theta$ for which $M(x) = 0$ is $\theta = 0$. Choose $c = 1/b$ noting that b is the slope of the response function. Then for any initial value, x_1, the expected value of x_2 can be easily computed since

$$x_2 = x_1 - \frac{1}{b}\{Y(x_1) - 0\}$$

implies that

$$E(x_2) = x_1 - \frac{1}{b} E\{Y(x_1)\},$$

where

$$E\{Y(x_1)\} = M(x_1) = bx_1.$$

Hence

$$E(x_2) = x_1 - \frac{1}{b} bx_1 = 0,$$

the desired θ.

Thus in the linear case the correct choice of c will move the estimate to the neighborhood of θ early in the process as evidenced by the fact that the first choice actually produces an unbiased estimate.

B. ESTIMATING THE SLOPE TO IMPROVE ASYMPTOTIC VARIANCE

It was noted by Wetherill [Ref. 130] that in the simple case where M(x) is a linear function that it can be shown that when we use as the sequence of a_n, $a_n = c/n$ that choice of c is critical to the efficiency of the process where efficiency is defined as the reciprocal of the ratio of the variance for a given c to the variance at $c = M'(\theta)$. See Table 1 (also see Hodges and Lehmann [Ref. 64]).

TABLE 1

Asymptotic Efficiency of the Robbins-Monro Process as a Function of $c/M'(\theta)$

$c/M'(\theta)$	0.50	0.75	1.00	1.25	1.50	2.00	2.50
efficiency	0	0.88	1.00	0.96	0.88	0.75	0.64

Table 1 shows that there is a large range of c for which the process is very efficient, with $c = M'(\theta)$ being optimal. It also would imply that it is better to overestimate the value of c than to underestimate.

Burkholder [Ref. 12] discussed the possibility of estimating the slope of M at θ but this procedure was not investigated further under Venter [Ref. 121] presented an extension of the Robbins-Monro procedure which estimates the slope of the regression function at the root. The method is similar to the Kiefer-Wolfowitz procedure in that at each step two observations are taken, namely $Y'=Y(x_n+c_n)$ and $Y'' = Y(x_n-c_n)$ where $c_n = cn^{-\gamma}(1 + o(1))$, $c > 0$, $0 < \gamma \leq \frac{1}{2}$. Venter required that we know constants a and b such that $0 < a < M'(\theta) < b < \infty$. At each step he estimated the slope by B_n where

$$B_n = n^{-1} \sum_{j=1}^{n} (y_j' - y_j'')/2c_j,$$

and then kept the estimated slope within the established bounds by using A_n as the estimate of the slope where

$$A_n = \begin{cases} a & \text{if } B_n < a \\ B_n & \text{otherwise} \\ b & \text{if } B_n > b \end{cases}$$

Venter then defined the recursive relation

$$x_{n+1} = x_n - d_n A_n^{-1} \tfrac{1}{2}(y_n' + y_n''),$$

where

$$d_n = n^{-1}(1 + O(n^{-\frac{1}{2}})).$$

Venter showed that if in the choice of $\{c_n\}$ that $\frac{1}{4} < \gamma < \frac{1}{2}$ then

$$n^{\frac{1}{2}}(x_n - \theta) \xrightarrow{\mathcal{L}} N(0,\sigma^2/2(M'(\theta))^2),$$

and

$$n^{\frac{1}{4}}(A_n - M'(\theta)) \xrightarrow{\mathcal{L}} N(0,\sigma^2/2(1 + 2\gamma)).$$

However if $\gamma = \frac{1}{4}$ then

$$n^{\frac{1}{2}}(x_n-\theta) \xrightarrow{\mathcal{L}} N(-2\alpha_2 c^2/M'(\theta),\sigma^2/2(M'(\theta))^2),$$

and

$$n^{\frac{1}{4}}(A_n-M'(\theta)) \xrightarrow{\mathcal{L}} N(0,\sigma^2/3c^2).$$

Venter stated that in the case of $\gamma < \frac{1}{4}$ the bias in the estimate, x_{n+1} of θ, will dominate the error. Therefore the choice of $\gamma = \frac{1}{4}$ gives a small negative bias but decreases the variance in the estimate of the slope.

One might ask whether this modified procedure is actually at a disadvantage since it requires two observations per step. Venter showed that after n steps (2n observations) its variance is still achieving the minimum value of the old Robbins-Monro procedure after 2n steps (2n observations). Venter also provided an estimate of σ^2 so that confidence intervals could be constructed for his procedure.

Fabian [Ref. 47] later provided a sophisticated proof of asymptotic normality of Venter's procedure and of a similar procedure applied to the Kiefer-Wolfowitz method.

C. SMALL SAMPLE THEORY

Considering the practical applications of using stochastic approximation in experiments where infinite quantities of test items may not be available, it is justifiable to ask how small sample realizations compare with asymptotic theory. For instance if an experimenter has less than say 50 animals with which to determine the LD_{50} (Lethal Dose 50%) then one may be concerned with designing a stochastic approximation method with which to obtain the "best" possible results and an estimate of the expected error.

1. Choice of x_1:

If one has prior information that θ (for say $M(\theta)= 0.50$) lies in a narrow interval and picks x_1 in that interval then one can expect the estimates to arrive in the neighborhood of θ within a few observations. If, however there is little prior knowledge of the magnitude of θ, then an initial bad choice of x_1 can induce a large bias term which will dominate the observations for many steps.

2. Choice of Multiplier, a_n:

As previously discussed $a_n = c/n$ where c equals the inverse of the slope of $M(\cdot)$ at θ is optimal in a sense. Thus one must accurately estimate c for optimal conditions.

If c is too small the step sizes may be too small to get to θ before the number of samples are depleted. Similarly if c is too large the estimate may overshoot θ back and forth. For a detailed analysis see Section V.A.

3. How to Allocate Samples:

If an experimenter has N samples to test, should he test one at each step and take N steps or test m at each step and take n = N/m steps? Note that taking more than one observation at each level, x_i, yields a more accurate estimate of $M(x_i) = E(Y|x_i)$. It was noted by Wetherill [Ref. 130] and by Cochran and Davis [Ref. 17], and was proven by Block [Ref. 5], that the variance of the estimate of θ depends only on the total samples, not on the sampling scheme; however the corresponding bias term, and hence mean squarred error, is affected by the scheme.

Cochran and Davis presented two graphs which illustrate their analysis, which is reproduced here. In their notation σ = the standard deviation of the observation, Y(x), at $x = \theta$. (which in general will be unknown to us). Also note the following terminology:

MSE: Mean Squarred Error;

C_0 : Optimal choice of coefficient, c;

m : # of samples taken at each level;

n : # of levels or steps,

where nm = N = Total number of samples.

Figure 2 Figure 3

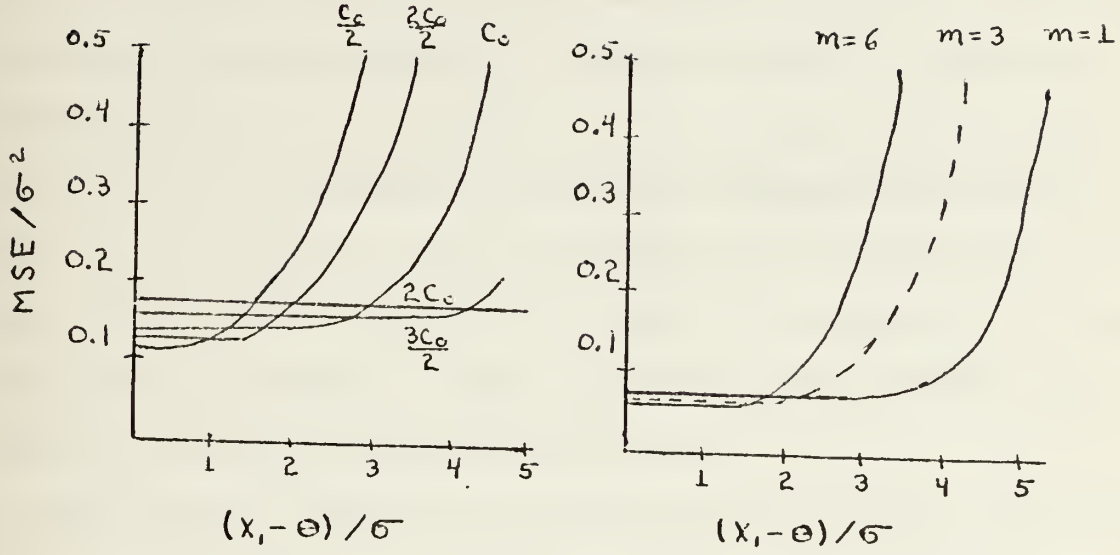


Figure 2 implies that if x_1 is relatively unknown that it is more desirable to overestimate c so we are not "trappe " by a large initial bias and small steps. Figure 3 implies that if the initial guess, x_1, is more than about 2σ away, then sampling should be done one at a time, while if the initial guess is very accurate, then the MSE's are smaller, although very slightly so, for larger m. Thus as a general rule unless we know that the initial guess is very accurate or unless the cost of setting up experiments at different levels is high, sampling should be conducted as one sample per level.

Another question which the experimenter may ask is how much more accurate an estimate becomes if he doubles the number of samples, say from scheme 1: m = 3, n = 8 to scheme 2: m = 6, n = 8. Doubling the value of m in this way reduces the variance of the estimate by almost exactly one-half, but produces only a slight decrease in the bias. Consequently if V, B, and M are the Variance, Bias and MSE for m = 3, n = 6 (scheme 1) then the corresponding MSE for m=6, n = 8 (scheme 2) can be predicted by the expression: $MSE_2 = (B^2 + V/2) = (B^2 + M)/2$. (This is assuming x_1 is the same for both schemes.) This expression overestimates the MSE, but at most by only a few percent.

For further results and comparisons of methods utilizing small sample theory see Cochran and Davis [Ref. 17], Davis [Ref. 22], Wetherill [Ref. 130], and Odell [Ref. 87].

D. ESTIMATION OF EXTREME QUANTILES

For estimates of quantiles near the mid-region of a quantal response curve the Robbins-Monro method appears to perform quite well. In fact for estimation of the $\theta_{.50}$ quantile both Wetherill [Ref. 130] and Davis [Ref. 22] showed that sample sizes as small as 35 produced results which were in good general agreement with asymptotic theory. However in areas away from the neighborhood of $\theta_{.50}$ the small sample estimates frequently have large biases and have variances greatly in excess of theoretical predictions. This behavior was also noted by Stillings and Logan [Ref. 108].

To try to explain this phenomenon Wetherill [Ref. 131] presented the following example.

Suppose an experimenter wishes to estimate $\theta_{.90}$ and that his initial level, x_1, is very close to the true value. Suppose further that the first observation is zero, a failure (as it will be about once every ten trials), then the second observation will be taken at the level

$$x_2 = x_1 - c(0 - 0.90) = x_1 + .90c.$$

This value, x_2, may well be far above $\theta_{.90}$. Assume that the next two values will be positive (a success). This leads to

$$x_3 = x_2 - \frac{c}{2}(1 - .90) = x_1 + .85c$$

and

$$x_4 = x_3 - \frac{c}{3}(1 - .90) = x_1 + .816c.$$

As can be easily observed the level of testing is very slowly returning to the vicinity of $\theta_{.90}$. In fact a minimum of about e^{10} observations are necessary to pass below x_1.

Methods using accelerated stochastic approximation tend to minimize this effect but the most interesting treatment of this area thus far has been done by Goodman, Lewis, and Robbins [Ref. 58]. Here a "maximum transformation" is

employed by taking multiple samples at a level. If it is desired to estimate $F(\theta) = .99$, where $F(x)$ is a cumulative density function, then V samples are taken at each level S. Here V is the solution of the equation

$$(0.99)^V = 0.50 .$$

Then let

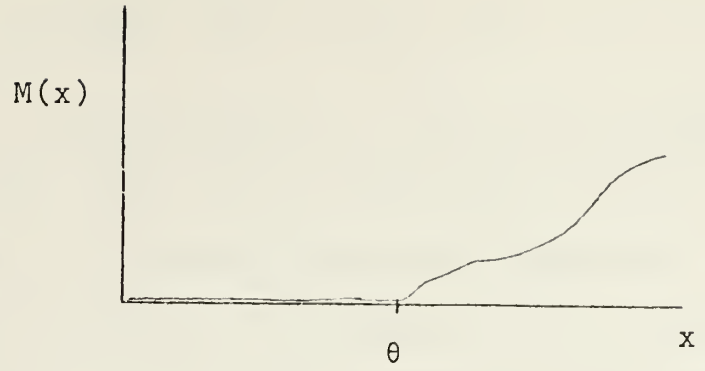
$$F_{max}(s) = \text{Prob}\{(\max_{1 \le i \le V} S_i) \le s\} = [F(s)]^V .$$

In this case the solution for V is V = 69, and 69 samples would be taken at each iteration. Thus the problem has been transformed into estimating the $\theta_{.50}$ level where the properties of the Robbins-Monro process are known to work well.

Yuguchi [Ref. 135] followed this same "maximum transformation" technique and then applied variance reduction and jack-knifing techniques to improve the rate of convergence and to reduce bias.

E. THE CASE WHERE M(x) STOPS BEING A CONSTANT

Consider a response function where there is no reaction for $x < \theta$. (I.e. $M(x) = 0$ for $x < \theta$ and $M(x) > 0$ for $x > \theta$.)



Often times one is interested in the level, θ, when response first occurs (see Guttman and Guttman [Ref. 60]). Friedman [Ref. 51] proved the following theorem.

THEOREM [Ref. 51]

Let the following conditions be satisfied;

(i) $|M(x)| \leq L|x| + K$;

(ii) $\sigma^2(x) \leq \sigma^2 < +\infty$;

(iii) if $x < \theta$, then $M(x) = 0$,

if $x > \theta$, then $M(x) > 0$;

(iv) for every $0 < \theta$ $\inf_{\delta \leq |x-\theta|} |M(x)| > 0$.

Then choose $\{a_n\}$, $\{d_n\}$ such that

$$a_n > 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty, \quad d_n > 0, \quad \sum_{n=1}^{\infty} a_n d_n = \infty,$$

$$d_n \to 0$$

and define the relation

$$x_{n+1} = x_n + a_n(d_n - y_n)$$

Then $x_n \to \theta$ with probability 1 and in mean square.

This theorem says that one can use stochastic approximation to find that point at which the regression function stops being a constant if the value of this constant is known. If one does not know the value of the constant, Friedman has proved another theorem which imposes sharper conditions on $M(x)$, for which x_n does converge to the desired value θ!

F. BOTH VARIABLES SUBJECT TO ERROR

In the usual Robbins-Monro procedure it is assumed that the regression function, $M(x_n)$ is observable subject to an error term, say v_x. One might ask under what conditions will the process converge if there exists a random error component, say u_x, in the level setting of x_n as in practice it is not always possible to precisely measure or set the desired amount. Dupac and Kral [Ref. 35] discussed two such cases. In the first case the error in setting the level is assumed to be unaffected by the experimenter. In the second case it is assumed that the error in the x level can be made arbitrarily small for an inversely proportional price. In this first case of "irreducible errors" Dupac and Kral proved the following theorem.

THEOREM [Ref. 35]

Assume that the following conditions are satisfied:

(i) $M(\cdot)$ is odd with respect to θ, i.e. $M(\theta + x) = -M(\theta - x)$ for all x;

(ii) $M(x)$ is strictly increasing;

(iii) $|M(x_2)-M(x_1)| \leq C_1 + C_2|x_2-x_1|$ for all x_1, x_2;

(iv) U_x is a symmetric random variable for each x, i.e. $P(U_x \leq c) = P(U_x > -c)$;

(v) Var $U_x \leq C_3$ for all x;

(vi) Var $V_x \leq C_4$ for all x.

Then the Robbins-Monro procedure defined by

$$x_{n+1} = x_n - a_n\{M(x_n + u_x) + v_x\}$$

converges to θ with probability 1 as well as in mean square.

In the second case of Dupac and Kral, where one can decrease the x setting errors, U_x, by an inversely proportional price, they proved what intuition would tell us was correct. They showed that it is needless to pay for high precision at the starting steps; the precision should be increased in the course of the approximation process.

G. THE CASE OF α UNKNOWN

Consider the following scenario: Suppose a scientist is comparing two drugs, a test drug and a control drug. He is interested in designing a biological assay to estimate

the number of dose units of the test drug necessary to elicit the same mean response as the standard dose of the control drug. Suppose further that the experimenter knows little about the shape of the response function associated with the test drug and about the probability distribution of response at any one dose level of either drug.

Make the following notational identifications: Let an observed response to the control drug administered at the standard dose level correspond to the random variable, Z, with mean α. Let the observed response to the test drug at dose level x correspond to Y(x) with mean, M(x). Let θ be the unknown dose level of the test drug such that $M(\theta) = \alpha$. Then under weak conditions on M(x), and the distributions of Y(x) and Z_i, the process defined by

$$x_{n+1} = x_n - a_n\{Y(x_n) - z_n\}$$

satisfies all known properties of the original Robbins-Monro procedure. It seems, as was noted by Hamilton [Ref. 62], that this procedure does not use all available information at each step. Since $n^{-1} \sum_{i=1}^{n} Z_i$ is a better estimator of α than just z_n, one would expect a smaller mean squarred error from the sequential estimate of α, especially in cases of small sample sizes. To analyze this Hamilton compared two processes.

Process 1 takes multiple observations at each step and computes an estimated value of α based only on the observations taken at that step (possibly only one).

Process 2 takes the same number of multiple observations at each step but computes the estimate of α based on all of the observations from the beginning of the process.

Hamilton then showed that under certain conditions it is better, in magnitude of mean squarred error, to take the most recent control observations (process 1) rather than taking sequential steps toward the mean of the control observation. This result, based on large sample theory, remains true in a simplified (linear) small sample situation.

VI. APPLICATIONS

In this Section several applications of stochastic approximation to a variety of fields will be presented. The first example is an application to a problem in biological research by Guttman and Guttman [Ref. 60]. It is initially presented since it is a simple, straightforward problem of the type for which the Robbins-Monro method was conceived (also see Hawkins [Ref. 63]). This straightforward use of the R-M method is also applicable to industrial process control as discussed by Comer [Refs. 18 and 19) where a lag in process response is incorporated into the formulation.

However, more practical use of stochastic approximation is based on the concepts of maximization or minimization of functions. Many problems which can be analytically solved if the response format is known fall nicely into the stochastic approximation framework since answers do not depen on the assumed parameterization. Also many problems based on a criterion, such as minimizing expected squarred error, can be computationally very difficult to solve, as the solutions may require matrix inversions, as in the multidimensional case. Many problems of this type (see Sardis, Nikolic and Fu [Ref. 99]) fall into the stochastic approximation framework and yield computationally simple algorithms which require very little storage space when performed on a digital computer.

In a recent book edited by Mendel and Fu [Ref. 83] a chapter has been devoted to applications of stochastic approximation methods. Also Tsypkin [Ref. 112] has nicely reviewed the important applicability of the Robbins-Monro process and related stochastic approximation methods to problems concerning pattern recognition, adaptive filters, adaptive automatic control systems, and adaption in operations research and reliability theory. Some of the additional papers which have considered these latter types of application of stochastic approximation are by Aizerman et al. [Ref. 1], Ernst [Ref. 38], Kailath and Schalkwijk [Ref. 67], Lee [Ref. 80], Sakrison [Refs. 93, 94], Sklansky [Ref. 106], Tsypkin [Ref. 111] and Ulrich [Ref. 116].

A. APPLICATION TO A PROBLEM IN BIOLOGICAL RESEARCH

Guttman and Guttman [Ref. 60] desired to treat Paramecium Caudatum cells with a substance, kinetin, which would stimulate cell division, and to estimate the time at which a certain level of this cell division was attained. They postulated that the ratio of the number of daily cell divisions of treated paramecia to untreated paramecia (K/C) was a monotone increasing function of time of exposure to kinetin. Guttman and Guttman stated that they had no idea of the underlying probability distribution concerning the ratio, K/C, thereby making stochastic approximation a very convenient scheme. A Robbins-Monro scheme was formulated to estimate the time at which K/C = 1.10. The initial

guess of X_1 = 30 hours was chosen with the expectation that the desired value of X was somewhere below this.

The sequence $\{a_n\}$ was chosen as 20/n to allow for large corrections in the first few steps and smaller corrections thereafter. The stochastic approximation sequence, as formulated for this problem, then looked like

$$X_{n+1} = X_n + \frac{20}{n} (1.10 - Y_n),$$

where Y_n = the observed response ratio at time X_n. Guttman and Guttman's table of observations, Y_n, and computed next levels, X_n, is reproduced in Table 2.

The experiment was terminated at n = 13 as no appreciable differences appeared among the X_n from trial 6 onward. Note that the mean value of the observations from n = 6 onwards is in fact equal to 1.10.

B. AN APPLICATION TO TAILORED TESTING

Suppose an educator or psychologist desires to measure some mental trait of an individual. For instance suppose it is desired to measure the level of difficulty of questions, X, such that the individual will get, say α = 70% of them correct. Suppose further that the educator has a bag full of questions, each assigned a level of difficulty, B_i, such that the probability that an individual, whose true ability is at level i, will correctly answer a question of difficulty B_i is equal to α = .70. This is similarly written

TABLE 2

Stochastic Approximation of Hours of Treatment Required with 1.5 mg/l Kinetin to Produce an Expected Ratio of Divisions Kinetin/Control Equal to 1.1.

TRIAL(n)	HOURS OF TREATMENT(x_n)	OBSERVED K/C(Y_n)	WEIGHT(a_n)
1	30	1.067	20
2	30.7	1.30	10
3	28.7	1.131	6.67
4	27.3	1.223	5
5	26.6	1.577	4
6	24.8	1.133	3.33
7	24.6	0.89	2.86
8	25.2	1.00	2.5
9	25.5	0.81	2.2
10	25.6	1.31	2
11	25.1	1.21	1.82
12	24.8	1.03	1.66
13	24.9	----	----

$$P_i(B_i) = \alpha.$$

This idea was presented by Lord [Ref. 82] who proposed a computer controlled testing scheme where questions of difficulty B_i would be recursively selected by the scheme

$$B_{i+1} = B_i + a_i\{Y(B_i) - \alpha\}.$$

Thus the scheme would eventually converge to the individual's true ability, provided that the assumptions were correct.

C. UPGRADING OF INERTIAL NAVIGATION SYSTEMS

Consider a navigational platform with several high grade Gyro's required for motion sensing. Bernard Lee [Ref. 80] suggested replacing all but one gyro with a lower grade, less expensive gyro. A supervisory system based on a continuous Keifer-Wolfowitz stochastic approximation algorithm similar to that developed by Sakrison [Ref. 90] is then used to estimate the drift rate of each of the low grade gyros and to apply a corrective signal. This concept permits each substandard gyro to acquire a precision approaching that of the higher gyro.

D. APPROXIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS

Consider the distribution $F(a) = \text{Prob}[X \leq a]$ where X is a scalar random variable. The problem is to find an

approximation to $F(\cdot)$ by a linear combination of a previously chosen vector of functions, $\underset{\sim}{\phi}^T(x) = (f_1(x), f_2(x),\ldots,f_n(x))$, where the superscript T denotes the transpose of the column vector $\underset{\sim}{\phi}(x)$. Thus we desire to find a column vector of coefficients, $\underset{\sim}{C}$, such that our approximation

$$\hat{F}(x) = \underset{\sim}{C}_F^T \underset{\sim}{\phi}(x)$$

minimizes some criterion such as minimizing expected squared error in a region of interest (a,b). Denote the mean square error as

$$J_F(\underset{\sim}{C}) = \int_a^b \{F(x) - \underset{\sim}{C}_F^T \underset{\sim}{\phi}(x)\}^2 dx.$$

Now minimizing $J_F(\underset{\sim}{C})$ is equivalent to solving the matrix equation

$$\frac{d(J_F(C))}{d\underset{\sim}{C}} = \int_a^b F(x) \underset{\sim}{\phi}(x)dx - \underset{\sim}{C}_F^T \int_a^b \underset{\sim}{\phi}(x) \underset{\sim}{\phi}^T(x)dx = 0$$

or

$$\int_a^b F(x) \underset{\sim}{\phi}(x)dx - K \underset{\sim}{C}_F^T = 0,$$

where

$$\underset{\sim}{K} = \int_a^b \underset{\sim}{\phi}(x) \underset{\sim}{\phi}^T(x)dx$$

is an n x n matrix.

Now define a random function $Z_F(y,x)$ such that

$$Z_F(y,x) = \begin{cases} 1 & \text{if} \quad y \leq x, \\ 0 & \text{if} \quad y > x, \end{cases}$$

and such that

$$E[Z_F(y,x)] = 1 \cdot F(x) + 0(1 - F(x)) = F(x).$$

Thus the regressive matrix equation

$$E \left\{ \int_a^b Z(y,x) \, \underset{\sim}{\phi}(x)dx \right\} - \underset{\sim}{K} \, \underset{\sim}{C}_F^T = 0$$

is equivalent to our previous equations for finding the minimum of the criterion, $J_F(\alpha)$. But this can now be solved by a stochastic approximation algorithm if successive independent samples of the random variable, Y, are availab e. The algorithm can be written as

$$\underset{\sim}{C}_F(j+1) = \underset{\sim}{C}_F(j) + a_j \, [\underset{\sim}{B}_F(Y(j) - \underset{\sim}{K} \, \underset{\sim}{C}_F(j)]$$

where we define

$$\underset{\sim}{B}_F(Y(j)) = \int_a^b Z_F(y(j),x) \underset{\sim}{\phi}(x)dx$$

i.e.

$$\underset{\sim}{B}_F(y(j)) = \begin{cases} \int_a^b \underset{\sim}{\phi}(x)dx & \text{if} \quad y(j) \leq a \\ \int_{y(j)}^b \underset{\sim}{\phi}(x)dx & \text{if} \quad a < y(j) < b \\ 0 & \text{if} \quad b \leq y(j), \end{cases}$$

where $y(j)$ is the J^{th} sample from the distribution and the sequence a_n satisfies

$$\sum_{J=1}^{\infty} a_J = \infty \qquad \text{and} \qquad \sum_{J=1}^{\infty} a_J{}^2 < +\infty.$$

Thus the above algorithm now fits the format of multi-dimensional stochastic approximation. In particular, if the matrix $\underset{\sim}{K}$ is positive definite, it satisfies the conditions of a theorem by Blum [Ref. 7, theorem 2].

Then the sequence $\underset{\sim}{C}_F(j)$ converges with probability 1 to the value which minimizes $J_F(\underset{\sim}{C})$. This value can be written as

$$\underset{\sim}{C}_F{}^* = K^{-1} \int_a^b F(x)\, \underset{\sim}{\phi}(x)dx,$$

but requires inversion of an n x n matrix to solve directly. Therefore the above algorithm enables one to find a minimum mean square error approximation to a distribution function for which the only available information is the collection

of sample values randomly selected. This algorithm is from a paper by Blaydon [Ref. 4] and can be similarly extended to approximate density functions.

A refinement of this algorithm by Blaydon was presented by Deuser and Lainiotis [Ref. 27]. The refinement incorporates a double stochastic approximation algorithm to recursively generate a matrix from each independent observation and then to recursively generate the estimate of the coefficient vector using the previously generated matrix as an observation. Deuser and Lainotis presented the example where the unknown probability is $F(x) = 1 - e^{-x}$ for $x \geq 0$.

The approximating function, $\hat{F}(x)$, is to be a weighted sum of the first three Laguerre polynomials

$$\underset{\sim}{\phi}(x) = \begin{cases} 1 \\ 1 - X \\ 1 - 2X + \frac{X^2}{2} \end{cases},$$

and the initial choice of the coefficient vector is the zero vector. It can be shown analytically that the optimal coefficients are:

$$\underset{\sim}{C}^T = (.480 \quad -.186 \quad -.239)$$

In a computer simulation using 1000 samples and using the step sequence $a_n = 1/n$, Deuser and Lainiotis obtained estimates for the coefficients which, on the average, did

not differ by more than .01 in absolute value from the optimal coefficients.

E. AN APPROACH TO PATTERN RECOGNITION USING STOCHASTIC APPROXIMATION TO MINIMIZE RISK

Consider a mixture from two samples where an observation which is drawn at random is of type 1 with unknown probability p, and is then of type 2 with probability 1 - p. It is desired to measure some quality of the samples, call it x, and apply a decision rule say

$$d(x,a)\begin{cases} = 1 & \text{if} \quad x \leq a \\ = 2 & \text{if} \quad x > a, \end{cases}$$

where $a^* = \theta$ is some unknown value which minimizes a risk function, $R(d(x,a))$ which we have chosen. Since the choice of a completely specifies the decision rule and risk function, denote them $d(a)$ and $R(a)$.

Now $R(a)$ can be viewed as a regression function. By this it is meant that there exists a random variable, Y, with conditional probability distribution function $F(Y|a)$ such that

$$R(a) = E(Y|a).$$

Such a random variable, Y, is defined as follows:

Let Y (given a) = C_{ij} if Z is an observation which actually is of type i and is classified by d(z,a) = type j.

In general $C_{ii} = 0$.

Then a simple one dimensional stochastic approximation scheme can demonstrate the solution process. Consider a test sample where it is not known of which population each item is a member. Then define the scheme

$$a_{n+1} = a_n - \frac{b_n}{2\, d_n} (Y^+ - Y^-),$$

where $Y^+ = C_{ij}$ if sample Z_{2n-1} is actually of type i and $d(Z_{2n-1}, a_n + d_n)$ = type j. $Y^- = C_{ij}$ if sample Z_{2n} is actually of type i and $d(Z_{2n}, a_n - d_n)$ = type j, where A_1 is chosen arbitrarily and the conditions

$$\sum_{n=1}^{\infty} b_n = \infty,$$

$$\lim_{n \to \infty} d_n = 0,$$

$$\sum_{n=1}^{\infty} (b_n/d_n)^2 < +\infty,$$

are satisfied. Note that the risk function must satisfy

$$\sup_{1/k < a-\theta < k} \overline{D}\, R(a) > 0$$

and

$$\inf_{1/k<a-\theta<k} \underline{D}\, R(a) < 0 \quad \text{for all integers } K,$$

where $\overline{D}\, R(a)$ = the limit superior of

$$\frac{R(a+h) - R(a)}{h} \quad \text{for } h \to 0$$

and $\underline{D}\, R(a)$ = the limit inferior of

$$\frac{R(a+h) - R(a)}{h} \quad \text{for } h \to 0.$$

Note that R(a) does not have to be differentiable at all a.

Then if the above conditions are satisfied, a_n converges in probability to θ and $\lim_{n\to\infty} [(a_n - \theta)^2] = 0$.

Then the decision rule which minimizes the risk function is

$$d(X,\theta) = \begin{cases} 1 & \text{if } X \leq \theta, \\ 2 & \text{if } X > \theta. \end{cases}$$

The above one-dimensional scheme was presented by Cooper [Ref. 20] who stated that the application to a K-dimensional scheme including noise could be performed by modifying the above procedure to the multidimensional case of Blum [Ref. 7].

It is noted that the above sample falls into the framework of Bayesian learning and decision rules. An excellent paper by Chien and Fu [Ref. 13] discusses Bayesian related

learning procedures which can be shown to be a special case of stochastic approximation algorithms and hence can be carried out in computationally simple schemes as the one just presented.

VII. AREAS FOR FURTHER STUDY

This Section is devoted to stating particular areas where further work may be of interest. These ideas have been noted as either not being discussed in the current literature or as having been analyzed when required conditions were not satisfied.

A. DEVIATIONS FROM THE LINEAR CASE

Section III.G discussed the estimate of expected squared error for the linear case and mentioned that other than a sampling experiment by Teichroew, very little analysis had been done. Work needs to be done in this area to determine limits of departures from linearity where linear results remain valid.

B. STOPPING RULES

Stopping rules not based on bounded confidence interva s utilizing aymptotic normality are almost nonexistent. Wha is needed is some nonparametric stopping rule based on say, number of changes of sign of $(x_n - x_{n-1})$. Many authors have noted that this is a virtually untouched area yet almost nothing has appeared in the literature.

C. POSSIBLE WEAKENING OF CONDITIONS ON a_n

In Comer's paper "Application of Stochastic Approximation to Process Control" [Ref. 19], an error in the formulation of a Robbins-Monro process yields interesting results.

Comer mistakenly used the step sequence $a_n = 1/(n)^{.5}$ in a simulation comparison. Note that this sequence does not satisfy the requirement $\sum_{n=1}^{\infty} (a_n^2) < \infty$. However his results when compared with the same simulation using $a_n = 1/n$, which does satisfy the necessary requirements, shows that the sequence, $a_n = 1/n^{.5}$ gives comparable if not superior results. The idea to explore is (1) Comer's simulation error, or (2), can the conditions on a_n actually be weakened in practice to obtain more desirable results.

D. REPEAT SIMULATION OF THE KIEFER-WOLFOWITZ PROCESS

In a previous simulation comparison study of Kiefer-Wolfowitz type methods, Springer [Ref. 107] used as a sequence of norming constants the sequence where $a_{n+1} = a_{n/2}$. He discussed the result of finding a small sample bias which, one should note, can be attributed to the fact that this sequence does not satisfy the assumption that $\sum_{n=1}^{\infty} a_n = \infty$. Perhaps a new simulation study using proper coefficients i in order.

E. MULTIDIMENSIONAL EXTENSION OF DUPAC AND KRAL's RESULTS

Dupac and Kral [Ref. 35] (see Sec. V.F) examined the Robbins-Monro one dimensional case where there are errors in setting the X-level. They cited conditions where $X_n \xrightarrow{P} \theta$ when these errors exist. They noted that errors of this type make the Kiefer-Wolfowitz procedure practially inapplicable to this type of analysis, but speculated that a generalization to the multidimensional case might be of interest.

LIST OF REFERENCES

1. Aizerman, M., Braverman, E., and Rozunder, L., "The Robbins-Monro Process and the Method of Potential Functions," Automation and Remote Control, v. 26, p. 1882-1885, 1965.

2. Albert, A. E., and Gardner, L. A. Jr., Stochastic Approximation and Nonlinear Regression, Riverside Press, Cambridge, 1967.

3. Bertram, J. E., "Control by Stochastic Adjustment," AIEE Applications and Industry Report, 1960.

4. Blaydon, C. C., "Approximation of Distribution and Density Functions," Proceedings of IEEE, v. 55, p. 231-232, 1967.

5. Block, H.D., "Estimates of Error for Two Modifications of the Robbins-Monro Stochastic Approximation Method," Annals of Mathematical Statistics, v. 28, p. 1003-1010, 1957.

6. Blum, J. R., "Approximation Methods Which Converge with Probability One," Annals of Mathematical Statistics, v. 25, p. 382-386, 1954.

7. Blum, J. R., "Multidimensional Stochastic Approximation Methods," Annals of Mathematical Statistics, v. 25, p. 737-744, 1954.

8. Blum, J. R., "A note on Stochastic Approximation," Proceedings of the American Math. Society, v. 9, p. 404-407, 1958.

9. Box, G. E. P., and Jenkins, G. M., "Some Statistical Aspects of Adaptive Optimization and Control," Journal of the Royal Statistical Society, Series B, v. 24, p. 297-343, 1962.

10. Box, G. E. P., and Wilson, K. B., "On the Experimental Attainment of Optimum Conditions, " Journal of the Royal Statistical Society, Series B, v. 13, p. 1-45, 1951.

11. Bunke, H., "Stochastic Approximation of the 50 Percent Dosage in Alternative Dosage-Effect Investigations," Biom. Zeit. v. 9, p. 236-239, 1967.

12. Burkholder, D. L., "On a Class of Stochastic Approximation Procedures," Annals of Mathematical Statistics, v. 27 p. 1044-1059, 1956.

13. Chien, Y. T., and Fu, K. S., "On Bayesian Learning and Stochastic Approximation," IEEE Transactions on Systems, Science and Cybernetics, V. SSC-3, p. 28-38, 1967.

14. Chung, K. L., "On Stochastic Approximation Methods," Annals of Mathematical Statistics, v. 25, p. 463-483, 1954.

15. Cochran, W. G., and Davis, M., "Sequential Experiments for Estimating the Mean Lethal Dose," Internationaux du Centre National de la Recherche Scientifique, v. 110 p. 181-194, 1963.

16. Cochran, W. G., and Davis, M., "Stochastic Approximation to the Median Lethal Dose in Bioassay," Stochastic Models in Medicine and Biology, University of Wisconsin Press, 1964.

17. Cochran, W. G., and Davis, M. "The Robbins-Monro Method for Estimating the Median Lethal Dose," Journal of the Royal Statistical Society, Series B, v. 27, p. 28-44, 1965.

18. Comer, J. P. Jr., "Some Stochastic Approximation Porcedures for Use in Process Control," Annals of Mathematical Statistics, v. 35, p. 1136-1146, 1964.

19. Comer, J. P. Jr., "Application of Stochastic Approximation to Process Control," Journal of the Royal Statistical Society, Series B, v. 27, p. 321-331, 1965.

20. Cooper, D. B., "Adaptive Pattern Recognition and Signal Detection using Stochastic Approximation," IEEE Transactions on Electronic Computers, v. EC-13, p. 306-307, 1964.

21. Davis, M., Comparison of Sequential Experiments for Estimating the Median Lethal Dose, Ph.D. Thesis, Harvard University, 1965.

22. Davis, M., "Comparison of Sequential Bioassays in Small Samples," Journal of the Royal Statistical Society, Series B, v. 33, p. 78-87, 1971.

23. deFigueiredo, R. J. P., and Dyer, L. W. Extensions of Discrete Stochastic Approximation with Dynamics and Applications to Nonlinear Filtering, Rice University, Systems Report No. 20-5, 1968.

24. Derman, C., "An Application of Chung's Lemma to the Kiefer-Wolfowitz Stochastic Approximation Procedure," Annals of Mathematical Statistics, v. 27, p. 532-536, 1956.

25. Derman, C., "Stochastic Approximation," Annals of Mathematical Statistics, v. 27, p. 879-886, 1956.

26. Derman, C., and Sacks, J., "On Dvoretzky's Stochastic Approximation Theorem," Annals of Mathematical Statistics, v. 30, p. 601-606, 1959.

27. Deuser, L. M., and Lainotis, D. G., "Minimum Mean Square Error Approximation of Unknown Probability Distribution Functions," IEEE Proceedings of the 7th Symposium on Adaptive Processes, UCLA, 1968.

28. Diaz, J.A., State Estimation for Nonlinear Systems, M. S. Thesis, U. S. Naval Postgraduate School, 1970.

29. Doob, J. L., Stochastic Processes, John Wiley, New York 1953.

30. Driml, M. and Hans, O. "Continuous Stochastic Approximations," Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Czechoslovak Academy of Sciences, Prague, p. 113-122, 1960.

31. Driml, M. and Nedoma, J., "Stochastic Approximations for Continuous Random Processes," Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Czechoslovak Academy of Sciences, Prague, p. 145-158, 1960.

32. Dupac, V. "On the Keifer-Wolfowitz Approximation Method," Selected Translations of Mathematical Statistics and Probability, v. 4, p. 43-69, 1957.

33. Dupac, V., "Note on Stochastic Approximation Methods," Czechoslovak Mathematical Journal, v. 8, p. 139-149, 1958.

34. Dupac, V., "A Dynamic Stochastic Approximation Method," Annals of Mathematical Statistics, v. 36, p. 1695-1702, 1965.

35. Dupac, V., and Kral, V., "Robbins-Monro procedure with both Variables Subject to Experimental Error," Annals of Mathematical Statistics, v. 43, p. 1089-1095, 1972.

36. Dvoretzky, A., "On Stochastic Approximation," Proceedings of The Third Berkeley Symposium on Mathematical Statistics and Probability, v. I, p. 39-55, 1956.

37. Eppling, M. L., A Multivariate Stochastic Approximation Procedure, Stanford University, Department of Statistics, Technical Report No. 5, 1964.

38. Ernst,R.,"Measurement and Control of Magnetic Field Homogeneity," Rev. Scient. Instrum., v. 39, p. 998-1012, 1968.

39. Fabian, V., "A Stochastic Approximation Method for Finding Optimal Conditions in Experimental Work and in Self-adapting Systems," Apl. Mat., v. 6, p. 162-183, 1961.

40. Fabian, V., "Stochastic Approximation Methods," Czechoslovak Mathematical Journal, v. 10, p. 123-159, 1960.

41. Fabian, V., "A New One Dimensional Stochastic Approximation Method for finding a local minimum of a function," Transactions of the Third Prague Conference on Information Theory, p. 85-105, 1964.

42. Fabian, V., "Stochastic Approximation of Constrained Minima," Transactions of the Fourth Prague Conference on Information Theory, p. 277-290, 1965.

43. Fabian, V., "A Survey of Deterministic and Stochastic Methods of Approximation for Minimization of Function ," presented at the Institute of Information and Automat on, Czechoslovak Academy of Sciences, Prague, June 1965.

44. Fabian, V., "On Asymptotic Normality in Stochastic Approximation," Michigan State University, Statistical Laboratory Publication, 1967.

45. Fabian, V., "Stochastic Approximation of Minima with Improved Asymptotic Speed," Annals of Mathematical Statistics, v. 38, p. 191-200, 1967.

46. Fabian, V., "On Choice of Design in Stochastic Approximation Methods," Annals of Mathematical Statistics, v. 39, p. 457-465, 1968.

47. Fabian, V., "On Asymptotic Normality in Stochastic Approximation," Annals of Mathematical Statistics, v.39, p. 1327-1332, 1968.

48. Fabian, V., "Stochastic Approximation for Smooth Functions," Annals of Mathematical Statistics, v. 40, p. 299-302, 1969.

49. Farrell, R. H., Sequentially Determined Bounded Length Confidence Intervals, Ph. D. Thesis, University of Illinois, 1959.

50. Farrell, R. H., "Bounded Length Confidence Intervals for the Zero of a Regression Function," Annals of Mathematical Statistics, v. 33, p. 237-247, 1962.

51. Friedman, S., "On Stochastic Approximations," Annals of Mathematical Statistics, v. 34, p. 343-346, 1963.

52. Fu, K. S., Nikolic, Z. Z., Chien, T. Y., and Wee, W. G., On The Stochastic Approximation and Related Learning Techniques, Purdue University, School of Electrical Engineering, Technical Report TR-EE-66-6, 1966.

53. Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1968.

54. Fu, K. S., "Learning Control Systems-Review and Outlook," IEEE Transactions on Automatic Control, v. AC-15, 1970.

55. Fu, K. S., (edited by) Pattern Recognition and Machine Learning, Plenum Press, New York, 1971.

56. Gardner, L. A. Jr., Stochastic Approximation and its Applications of Prediction and Control Synthesis, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, p. 241-258, Academ c Press, New York, 1963.

57. Gladyshev, E. G., "On Stochastic Approximation," Teo ya Veroyatnostei i ee Primenenie, v. 10, 1965.

58. Goodman, A. S., Lewis, P. A. W., and Robbins, H. E., Simultaneous Estimation of Large Numbers of Extreme Quantiles in Simulation Experiments, Communications in Stastics, v. 2, 1973 (to appear).

59. Gray, K. B., "The Application of Stochastic Approximation to the Optimization of Random Circuits," American Mathematical Society, Proceedings of Symposium in Applied Mathematics, V. 16, p. 178-192, 1962.

60. Guttman, L. and Guttman, R., "An Illustration of the Use of Stochastic Approximation," Biometrics, v. 15, p. 551-559, 1959.

61. Hans, A., and Spacek, A., "Random Fixed Point Approximation by Differential Trajectories," Transactions of the Second Prague Conference on Information Theory, Statistical Decisions and Random Processes, p. 203-213,1960.

62. Hamilton, M. A., "The Stochastic Approximation Approach to a Discrimination Problem," Annals of Mathematical Statistics, v. 43, p. 1096-1109, 1972.

63. Hawkins, D. F., "Observations on the Application of the Robbins-Monro Process to Sequential Toxicity Assays," British Journal of Pharmac, Chemother, v. 22, p. 392-402, 1964.

64. Hodges, J. L., and Lehmann, E. L., "Two Approximations to the Robbins-Monro Process," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, p. 95-104, 1956.

65. Holmes, J. K., "Two Stochastic Approximation Procedures for Identifying Linear Systems," IEEE Transactions on Automatic Control, v. AC-14, p. 292-295, 1969.

66. Johnson, M. A., "A Note on the Interpretation of B. Springer's Simulation Studies of the Kiefer-Wolfowitz Process," Biometrika, v. 57, p. 673-674, 1970.

67. Kailath, T., and Schalkwijk, J., "A Coding Scheme for Additive Noise Channels with Feedback," IEEE Transactions on Information Theory, v. IT-12, p. 172-182, 1966.

68. Kallianpur, G., "A Note on the Robbins-Monro Stochastic Approximation Methods," Annals of Mathematical Statistics, v. 25, p. 386-388, 1954.

69. Kesten, H., "Accelerated Stochastic Approximation," Annals of Mathematical Statistics, v. 29, p. 41-59, 1958.

70. Kiefer, J. and Wolfowitz, J., "Stochastic Estimation of the Maximum of a Regression Function," Annals of Mathematical Statistics, v. 23, p. 462-466, 1952.

71. Kitagawa, T., "Successive Processes of Statistical Controls," Mem. Fac. Sci. A. Kyushu. v. 13, p. 1-16, 1959.

72. Krasulina, T. P., "On Stochastic Approximation," Proceedings of the Fourth All-Union Congress on Probability Theory and Statistics, Gospolitnauchizdat vil'nyus, 1962.

73. Krasulina, T. P., "A Note on Some Stochastic Approximation Processes," Theory of Probability and its Applications, v. 7, p. 108-113, 1962.

74. Krasulina, T. P., "Remarks on Certain Stochastic Approximation Processes," Teoriya Veroyatnostei i ee Primenenie, v. 6, 1962.

75. Kullback, S., Information Theory and Statistics, John Wiley, New York, 1959.

76. Kushner, H. J., Adaptive Techniques for the Optimization of Binary Detection Systems, MIT, Lincoln Laboratory Report 22G-2, 9 November 1962.

77. Kushner, H. J., Methods for the Adaptive Optimization of Binary Detection Systems, MIT, Lincoln Laboratories Report 22G-2, 9 November 1962.

78. Kushner, H. J., "Hill Climbing Methods for the Optimization of Multiparameter Noise Disturbed Systems," Journal of Basic Engineering, p. 157-164, 1963.

79. Kushner, H. J., "A Note on the Maximum Sample Excursions of Stochastic Approximation Processes," Annals of Mathematical Statistics, v. 37, p. 513-516, 1966.

80. Lee, B., "Autonomous Upgrading of Inertial Navigation Systems," IEEE Transactions on Aerospace and Electronic Systems, v. 3, p. 623-636, 1967.

81. Loginov, N., "Survey: Methods of Stochastic Approximation," Automation and Remote Control, v. 27, p. 707-728, 1966.

82. Lord, F. M., Tailored Testing, An Application of Stochastic Approximation, Office of Naval Research Report No. NR-150-303, 1971.

83. Mendel, J. M., and Fu, K. S., Adaptive Learning and Pattern Recognition, Academic Press, New York, 1970.

84. Molverton, C. T., and Rawgen, J. T., "A Counter Example to Dvoretzky's Stochastic Approximation Theorem," IEEE Transactions on Information Theory, v. 14, p. 157-158, 1968.

85. Morozan, T., "On Stochastic Approximation," C. R. Acad. Sci., Paris A, v. 264, p. 633-635, 1967.

86. Newbold, P. M., A Stochastic Approximation Scheme with Accelerated Convergence Properties, Harvard University, Division of Engineering and Applied Physics, Technical Report No. 545, 1967.

87. Odell, P. L., An Empirical Study of Three Stochastic Approximation Techniques, NAVWEPS Report No. 7837, 1961.

88. Robbins, H. and Monro, S., "A Stochastic Approximation Method," Annals of Mathematical Statistics, v. 22, p. 400-407, 1951.

89. Rozov, Y. L., Chelpanov, I. B., "The Use of the Stochastic Approximation Method When the Possibility Arises of Errors with the 'Miss' Type," Automation and Remote Control, v. 29, 1968.

90. Sakrison, D. J., Application of Stochastic Approximation Methods to System Optimization, MIT Research Laboratory of Electronics Technical Report 391, 1962.

91. Sakrison, D. J., "Iterative Design of Optimum Filters for Non-mean Square Error Criteria," Transactions of the IEEE Professional Group on Information Theory, v. 3, p. 161-167, 1963.

92. Sakrison, D. J., "A Continuous Kiefer-Wolfowitz Procedure for Random Processes," Annals of Mathematical Statistics, v. 35, p. 590-599, 1964.

93. Sakrison, D. J., "Efficient Recursive Estimation; Application to Estimating the Parameters of a Covariance Function," Journal of Eng. Sci., v. 3, p. 461-483, 1965.

94. Sakrison, D. J., "Efficient Recursive Estimation of the Parameters of a Radar or Radio Astronomy Target," IEEE Transactions on Information Theory, v. 12, p. 35-41, 1966.

95. Sacks, J., "Asymptotic Distribution of Stochastic Approximation Procedures," Annals of Mathematical Statistics, v. 29, p. 373-405, 1958.

96. Sage, A. P., "Stochastic Approximation Algorithms for Specific Suboptimal Estimation and Control," Proceedings of the Third International Conference on System Sciences, Part 2, 1970.

97. Sardis, G. N., and Stein, G., "Stochastic Approximation Algorithms for Linear Discrete-time System Identification," IEEE Transactions on Automatic Control, v. AC-13, p. 515-523, 1968.

98. Sardis, G. N. and Stein, G., "A New Algorithm for Linear System Identification," IEEE Transactions on Automatic Control, v. AC-13, p. 592-595, 1968.

99. Sardis, G. N., Nikolic, Z. Z., and Fu, K. S., "Stochastic Approximation Algorithms for System Identification, Estimation, and Decomposition of Mixtures," IEEE Transactions on Systems, Science, and Cybernetics, v. SSC-5, p. 8-15, 1969.

100. Schmetterer, L., "Sur l'Iteration Stochastique," Le Calcul des Probabilites et Ses Applications, v. 87, p. 55-63, 1958.

101. Schmetterer, L., "Stochastic Approximation," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, v. I, p. 587-609, 1960.

102. Schmetterer, L., "On A Problem of Stochastic Approximation and A Matrix Summation Method For Sequences of Numbers," Metrika, v. 14, p. 273-276, 1969.

103. Seilken, R. L. Jr., Sequentially Determined Bounded Length Confidence Intervals for Stochastic Approximation Procedures of The Robbins-Monro Type, Ph.D. Dissertation, Florida State University, 1971.

104. Seilken, R. L. Jr., Some Stopping Rules for Stochastic Approximation Procedures, Florida State University Statistics Report M217, 1971.

105. Sinha, N. K., and Griscik, M. P., "A Stochastic Approximation Method," IEEE Transactions on Systems, Man and Cybernetics, v. SMC-1, p. 338-344, 1971.

106. Sklansky, J., "Learning Systems for Automatic Contro ," IEEE Transactions on Automatic Control, v. AC-11, p. 6-19, 1966.

107. Springer, B., "Numerical Optimization in the Presenc of Random Variability, The Single Factor Case," Biometrika, v. 56, p. 65-74, 1969.

108. Stillings, T. J., and Logan, R. S., A Computational Analysis of the Robbins-Monro Stochastic Approximation Method, M.S. Thesis, U. S. Naval Postgraduate School, 1965.

109. Teichroew, D., An Empirical Investigation of the Stochastic Approximation Method of Robbins and Monro, Working Paper of the National Bureau of Standards, Washington, D. C., 1953.

110. Tratenser, M., Application of Stochastic Approximation Methods to Field Artillery Precision Fire, M.S. Thesis, U. S. Naval Postgraduate School, 1973.

111. Tsypkin, Y., "Establishing Characteristics of a Functioning Transformer from Randomly Observed Points," Automation and Remote Control, v. 26, p. 1878-1881, 1965.

112. Tsypkin, Y., "Adaptation, Training and Self-Organization in Automatic Systems," Automation and Remote Control, v. 27, p. 16-51, 1966.

113. Tsypkin, Y., "Use of the Stochastic Approximation Method in Estimating Unknown Distribution Densities from Observations, " Automation and Remote Control, v. 27, p. 432-434, 1966.

114. Tsypkin, Y., "Learning Control Systems," Automatika i Telemekhanika, v. 4, 1970.

115. Tsypkin, Y., Adaptation and Learning in Automatic Systems, Academic Press, New York, 1971.

116. Ulrich, M., "Some Probability Methods of Automatic Control," Automation and Remote Control, v. 25, p. 64-68, 1964.

117. Vaysboro, E. M., and Yudin, D. B., "Multiextremal Stochastic Approximation," Engineering Cybernetics, v. 6, p. 1-11, 1968.

118. Venter, J. H., and Gastwirth, J. L., Adaptive Statistical Procedures in Reliability and Maintenance Problems, Stanford University, Department of Statistics Technical Report No. 99, 1964.

119. Venter, J. H., On Stochastic Approximation Methods, Ph.D. Thesis, University of Chicago, 1963.

120. Venter, J. H., "On Dvoretzky Stochastic Approximation Theorems," Annals of Mathematical Statistics, v. 37, p. 1534-1544, 1966.

121. Venter, J. H., "An Extension of the Robbins-Monro Procedure," Annals of Mathematical Statistics, v. 38, p. 181-190, 1967.

122. Venter, J. H., "On Convergence of the Keifer-Wolfowitz Approximation Procedures," Annals of Mathematical Statistics, v. 38, p. 1031-1036, 1967.

123. Vosikova, M., Asymptoticke Viastnosti Kiefer-Wolfowitzovy Approximacni Metody, Ph.D. Thesis, Charles University, Prague, 1964.

124. Wasan, M. T., "A Technique of Stochastic Approximation," Proceedings of the Functional Analysis Symposium, QPPAM-10, p. 242-247, 1967.

125. Wasan, M. T., "On Stochastic Approximation Methods," abstract, Annals of Mathematical Statistics, v. 36, p. 1078, 1965.

126. Wasan, M. T., "Stochastic Approximation and its Application," American Mathematical Society Summer Seminar on Mathematics of Decision Sciences, 1967.

127. Wasan, M. T., "Sequential Optimum Procedures for Unbiased Estimation of A Binomial Parameter," Technometrics, v. 6, p. 259-272, 1964.

128. Wasan, M. T., Parametric Estimation, McGraw Hill, New York, 1969.

129. Wasan, M. T., Stochastic Approximation, Cambridge University Press, Cambridge, 1969.

130. Wetherill , G. B., "Sequential Estimation of Quantal Response Curves," Journal of the Royal Statistical Society, Series B, v. 25, p. 1-48, 1963.

131. Wetherill , G. B., Sequential Methods in Statistics, Methuen, London, 1966.

132. Wolfowitz, J., "On the Stochastic Approximation Method of Robbins and Monro," Annals of Mathematical Statistics, v. 23, p. 457-461, 1952.

133. Wolfowitz, J., "On Stochastic Approximation Methods," Annals of Mathematical Statistics, v. 27, p. 1151-1156, 1956.

134. Wolverton, C. T., "A Note on The Application of Dvoretzky's Theorem to Nonlinear Filtering," IEEE Transactions on Automatic Control, v. AC-14, p. 117, 1969.

135. Yuguchi, T., Quantile Estimation Using the Stochastic Approximation with Maximum Transformation and Jacknife, M.S. Thesis, U. S. Naval Postgraduate School, 1973.

136. Komlos, J., and Revesz, P., "On the Rate of Convergence of the Robbins-Monro Method," Zeitschrift fur Wahrscheinlichkeitstheorie, v. 25, p. 39-47, 1972.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Asst Prof Bruno Shubert, Code 55Sy Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	5
4.	Asst Prof P. A. W. Lewis, Code 55Le Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
5.	Asst Prof M. U. Thomas, Code 55 To Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
6.	Assoc Prof D. R. Barr, Code 55Bn Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
7.	Library, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	30
8.	LCDR William H. Wright IV, USN COMCARDIV FIVE Staff FPO, San Francisco, California 96601	1
9.	LT Thomas K. Scheber, USN 246 Beach Road Marina, California 93933	5

10. Chief of Naval Personnel 1
Pers 11b
Department of the Navy
Washington, D.C. 20370

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

ORIGINATING ACTIVITY (Corporate author)

Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

REPORT TITLE

Stochastic Approximation: A Survey

DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Master's Thesis; March 1973

AUTHOR(S) (First name, middle initial, last name)

homas Keith Scheber

REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1973	104	136

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

pproved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

Since 1951 when Robbins and Monro's pioneering paper on stochastic pproximation was published, many articles have appeared dealing with xtensions, modifications, methods and applications of stochastic pproximation. While the concepts involved are relatively simple, ut mathematically difficult, the information concerning specific esults has been widely scattered and difficult to collect for the interested researcher. This paper will attempt to discuss the major esults and will provide the necessary references to direct the user o more specific findings.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

A-31408

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
stochastic approximation						
sequential estimation						

Thesis
S285 Scheber
c.1 Stochastic approximation: a survey.

144022